Level-based analysis of genetic algorithms and other search processes

Corus, Dogan; Dang, Duc-Cuong ; Eremeev, Anton V.; Lehre, Per Kristian

DOI:
10.1109/TEVC.2017.2753538

Level-Based Analysis of Genetic Algorithms and Other Search Processes

Dogan Corus[1], Duc-Cuong Dang[2], Anton V. Eremeev[3], and Per Kristian Lehre[4]

[1]University of Sheffield, United Kingdom
[2]University of Nottingham, United Kingdom
[3]Omsk Branch of Sobolev Institute of Mathematics, Russia
[4]University of Birmingham, United Kingdom

August 10, 2017

Abstract

Understanding how the time complexity of evolutionary algorithms (EAs) depend on their parameter settings and characteristics of fitness landscapes is a fundamental problem in evolutionary computation. Most rigorous results were derived using a handful of key analytic techniques, including drift analysis. However, since few of these techniques apply effortlessly to population-based EAs, most time complexity results concern simple EAs, such as the (1+1) EA.

We present the *level-based theorem*, a new technique tailored to population-based processes. It applies to any non-elitist process where offspring are sampled independently from a distribution depending only on the current population. Given conditions on this distribution, our technique provides upper bounds on the expected time until the process reaches a target state.

The technique is demonstrated on pseudo-Boolean functions, the sorting problem, and approximation of optimal solutions in combinatorial optimisation. The conditions of the theorem are often straightforward to verify, even for Genetic Algorithms and Estimation of Distribution Algorithms which were considered highly non-trivial to analyse. The proofs for the example applications are available in the supplementary materials. Finally, we prove that the theorem is nearly optimal for the processes considered: Given the information the theorem requires about the process, a much tighter bound cannot be proved.

1 Introduction

The theoretical understanding of Evolutionary Algorithms (EAs) has advanced significantly over the last decades. Significant progress in developing and understanding a formal model of canonical GAs and their generalisations was made in the nineties using dynamical systems [54]. Notably, the behaviour of the dynamical systems model is closely related to the local optima structure of the

problem in the case of binary search spaces [55]. However, most of these findings relate to the infinite population limit, from which it is difficult to derive statements about performance. Researchers from theoretical computer science argued in the early 2000s that EC theory had attempted to make either too general, or too precise statements [3]. Instead, one should develop techniques for deriving rigorous statements about worst-case optimisation time, starting from the simplest possible settings. Much of the work assumed a population size of one and no crossover operator, e.g., the (1+1) EA [20].

Early analyses of EAs with larger population sizes often ignored the recombination operator. The family tree technique was introduced in [58] to analyse the $(\mu+1)$ EA. The performance of the $(\mu+\mu)$ EA for different settings of the population size was analysed in [30], using Markov chains to model the search processes, and in [4], using a similar argument to fitness levels. The analysis of parallel EAs in [35] also made use of the fitness levels argument. The inefficiency of standard fitness proportionate selection without scaling was shown in [44] and in [36], using drift analysis [28]. In the recently introduced switch analysis, the progress of the EA is analysed relative to a reference process that is easier to understand [60].

Runtime analyses taking into account recombination often aimed at understanding how evolutionary search can benefit from sexual reproduction in specific settings. For the simple linear OneMax problem, it is known that the Standard Genetic Algorithm (SGA) [54] is inefficient, even when crossover is enabled [45]. For a variant of the $(\mu+1)$ EA, crossover can speed up by a factor of 2 compared to the (1+1) EA [51]. GAs with even higher (non-constant) speedups on OneMax are known, but they rely on non-conventional reproduction mechanisms [16,17]. The unrestricted black-box complexity of OneMax is $\Omega(n/\log(n))$ [21], the speedup of any unrestricted black-box algorithm relative to the (1+1) EA is therefore at most $\mathcal{O}\left(\log(n)^2\right)$. More complex settings are required to show further speedups. So-called convex search algorithms, which include non-elitist GAs with gene pool recombination and no mutation, have been analysed on *quasi-concave fitness landscapes*. [41]. As a special case, convex search algorithm has expected runtime $\mathcal{O}(n \log n)$ on LeadingOnes, i.e., a speedup of $\Theta(n/\log(n))$ compared with the (1+1) EA. [47] considered non-elitist GAs, also without mutation, on noisy OneMax. The $(\mu+1)$ GA decreases the runtime on the Jump problem, however this was first only shown for artificially small crossover probabilities [31, 32]. For realistic crossover probabilities, the $(\mu+1)$ GA decreases the runtime by an exponential factor on instances of an FSM testing problem, however, this result assumes a deterministic crowding diversity mechanism [39]. It was recently shown that the standard $(\mu+1)$ GA without any modification has a speed up of $\Omega(n/\log(n))$ on the Jump problem compared to the mutation-only variant $(\mu+1)$ EA [10].

Estimation of Distribution Algorithms (EDA), a relatively new type of EAs [34], build explicit probabilistic models in every generation from the fittest individuals, from which the next generation is sampled. The sequence of probability distributions should converge to a distribution concentrated on the optimal search points. Traditional EAs can be seen as special cases of EDAs, where the probability distributions are given implicitly via their genetic operators and selection mechanisms. Most theoretical studies of EDAs have considered convergence and scalability properties [27,43,46,50,61], and rigorous runtime analyses of EDAs are still rare. Droste's $\Theta(Kn)$ bound on the expected runtime on linear

functions for cGA with update parameter $K \geq n^{1+\varepsilon}$, is an early rigorous result on the runtime of EDAs [19]. Recent runtime analyses have demonstrated the noise robustness of cGA [25], as well as the impact of its update parameter [52]. The runtime of UMDA, a more complex EDA [42], has been analysed in a series of papers [5–8, 13, 37, 59]. Chen et al. describe easy and hard functions for the UMDA under the so-called "no-random-error" assumption and with a sufficiently large population [6]. This assumption was lifted in [8], but the analysis still assumed an unrealistically large population size, leading to too highs bound on the expected runtime. It is usually necessary to impose margins on the probability distribution of the UMDA [7], however, the only known setting where the UMDA outperforms the (1+1) EA assumes the UMDA without margins [5]. An early variant of the level-based method provided the first upper bound of $\mathcal{O}(n\lambda \log \lambda)$ on the runtime of UMDA with realistic population size $\lambda = \Omega(\log n)$ on OneMax [13]. Recently, a more precise application of the level-based method tightened this bound to $\mathcal{O}(n\lambda)$ under the assumption that the parent population size is $\mu = \mathcal{O}(\sqrt{n})$ [37]. For offspring population sizes $\lambda = \mathcal{O}(\log n)$, this runtime bound is tight, because it matches the lower bound of $\Omega(\mu n + n \log n)$ shown via drift analysis [33]. A different argument than the level-based method yielded a similar upper bound [59] for the UMDA.

This paper shows that all non-elitist EAs with or without crossover, and even EDAs, can be cast and analysed in the same framework. An early version of the paper was communicated in [9]. This followed from work dating back to the introduction of a fitness level technique for *non-elitist* EAs with linear ranking selection [40], later on generalised to many selection mechanisms and *unary variation operators* [36], with a refined result in [14]. The original fitness level technique and its generalisation to the level-based technique have already found several applications, including analysis of EAs in uncertain environments, such as partial information [14], noisy fitness functions [12], and dynamic fitness functions [11]. It has also been applied to analyse the runtime of complex algorithms such as self-adaptive EAs [15], pointing out multi-modal fitness landscapes where they outperform classical, elitist EAs.

The present work improves the main result of [9] in many aspects. A more careful analysis of the population dynamics leads to a much tighter expression of the runtime bound compared to [9]. In fact, we prove that the new bounds are close to optimal for the class of search procesess the theorem applies to. This immediately implies improved results in the previously mentioned applications. In particular, the leading term in the runtime is improved by a factor of $\Omega(\delta^{-3})$, where δ characterises how fast good individuals can populate the population. This significantly improves the results of [12] and [14] concerning noisy optimisation, for which δ is often very small (e. g., $1/n$). We also recommend a stepwise guideline for how to apply the theorem to new settings. Example applications are given for the cases of GAs and UMDA in optimising standard pseudo-Boolean functions, a simple combinatorial problem, and in searching for local optima of NP-hard problems.

This paper has the following structure: Section 2 describes the class of algorithms covered by the level-based theorem, showing that many GAs and EDAs are special cases. The section then states the main theorem and corollaries for special cases, followed by their proofs. Sections 4 and 5 apply the level-based theorem to the Simple Genetic Algorithm (SGA) and UMDA on example problems. Section 6 proves that the level-based theorem is tight for the class

of algorithms considered. Section 7 concludes the paper. Some proofs have been omitted due to space restrictions, but are available in the supplementary materials.

2 Main result

2.1 Abstract algorithmic scheme

We consider population-based processes as stochastic processes $(P_t)_{t\in\mathbb{N}}$, where for each "generation" $t \in \mathbb{N}$, $P_t = (x_1, \ldots, x_\lambda) \in \mathcal{X}^\lambda$ is a vector of λ *individuals*, and where the set $\mathcal{X}$ represents the "search space" or "genospace". Our goal is to estimate the expected number of generations until the population contains at least one element in some given subset of $\mathcal{X}$. Our main assumption is that for every generation $t \in \mathbb{N}$, each individual in generation P_{t+1} is obtained by independent sampling from a distribution $D(P_t)$. Intuitively, D describes the randomised process determining how new individuals are produced, and may include fitness evalations, selection, variation operators, external noise etc. Formally, D is a mapping from the set of all possible populations $\mathcal{X}^\lambda$ into the space of probability distributions over $\mathcal{X}$. This scheme is summarised in Algorithm 1.

Algorithm 1 Population-based algorithm.

Require:
A finite state space $\mathcal{X}$, and population size $\lambda \in \mathbb{N}$,
a mapping D from $\mathcal{X}^\lambda$ to the space of prob. dist. over $\mathcal{X}$,
and an initial population $P_0 \in \mathcal{X}^\lambda$.

1: **for** $t = 0, 1, 2, \ldots$ until termination condition met **do**
2: Sample $P_{t+1}(i) \sim D(P_t)$ independently for all $i \in [\lambda]$.
3: **end for**

Algorithm 1 covers many non-elitist search heuristics, such as Stochastic Beam Search [54], Estimation of Distribution Algorithms, and Genetic Algorithms [26].

For example, the GA given by Algorithm 2 is a special case of Algorithm 1, where the operator D corresponds to lines 3–5 in Algorithm 2. Here, the random operator `select`: $\mathcal{X}^\lambda \to [\lambda]$ represents a selection mechanism, which given a vector of λ individuals, returns the index of the individual to be selected. The selection mechanism is typically defined relative to a fitness function $f\colon \mathcal{X} \to \mathbb{R}$. The GA uses the two variation operators `mutate`: $\mathcal{X} \to \mathcal{X}$, and `crossover`: $\mathcal{X} \times \mathcal{X} \to \mathcal{X}$.

Algorithm 2 Genetic Algorithm (GA)

Require:
A finite set $\mathcal{X}$, a population size $\lambda \in \mathbb{N}$, a recombination rate $p_c \in (0,1]$, and an initial population $P_0 \in \mathrm{Unif}(\mathcal{X}^\lambda)$.

1: **for** $t = 0, 1, 2, \ldots$ until termination condition met **do**
2: **for** $i = 1$ to λ **do**
3: $u := P_t(\texttt{select}(P_t))$, $v := P_t(\texttt{select}(P_t))$.
4: With prob. $p_c, x := \texttt{crossover}(u, v)$ **else** $x := u$.
5: $P_{t+1}(i) := \texttt{mutate}(x)$.
6: **end for**
7: **end for**

2.2 The level-based theorem

We now state the main result of the paper: a general technique for obtaining upper bounds on the expected runtime of any process that can be described in the form of Algorithm 1. We use the following notation. The natural logarithm is denoted by $\ln(\cdot)$, and $[n] = \{1, \ldots, n\}$ denotes the first n natural numbers. Suppose that for some m there is an ordered partition of $\mathcal{X}$ into subsets $(A_1, \ldots, A_m)$ which we call *levels*. We define $A_{\geq j} := \bigcup_{i=j}^{m} A_i$, i. e., the union of all levels above level j. The *canonical* partition of $\mathcal{X}$ with respect to a fitness function $f\colon \mathcal{X} \to \mathbb{R}$, is $x, y \in A_j$ if and only if $f(x) = f(y)$ (see, e. g., [36]). The partition is called f-based if $f(x) < f(y)$ for all $x \in A_j$, $y \in A_{j+1}$ and all $j \in [m-1]$. As a result of the algorithmic abstraction, our main theorem is not limited to this particular type of partition. Let $P \in \mathcal{X}^\lambda$ be a population vector of a finite number $\lambda \in \mathbb{N}$ of individuals. Given any subset $A \subseteq \mathcal{X}$, we define $|P \cap A| := |\{i \mid P(i) \in A\}|$, i. e. the number of individuals in P that belong to A.

Theorem 1. *Given a partition $(A_1, \ldots, A_m)$ of $\mathcal{X}$, define $T := \min\{t\lambda \mid |P_t \cap A_m| > 0\}$, where for all $t \in \mathbb{N}$, $P_t \in \mathcal{X}^\lambda$ is the population of Algorithm 1 in generation t. If there exist $z_1, \ldots, z_{m-1}, \delta \in (0,1]$, and $\gamma_0 \in (0,1)$ such that for any population $P \in \mathcal{X}^\lambda$,*

(G1) *for each level $j \in [m-1]$, if $|P \cap A_{\geq j}| \geq \gamma_0 \lambda$, then*

$$\Pr_{y \sim D(P)}(y \in A_{\geq j+1}) \geq z_j,$$

(G2) *for each level $j \in [m-2]$, and all $\gamma \in (0, \gamma_0]$ if $|P \cap A_{\geq j}| \geq \gamma_0 \lambda$ and $|P \cap A_{\geq j+1}| \geq \gamma \lambda$, then*

$$\Pr_{y \sim D(P)}(y \in A_{\geq j+1}) \geq (1+\delta)\gamma,$$

(G3) *and the population size $\lambda \in \mathbb{N}$ satisfies*

$$\lambda \geq \left(\frac{4}{\gamma_0 \delta^2}\right) \ln\left(\frac{128m}{z_* \delta^2}\right), \text{ where } z_* := \min_{j \in [m-1]}\{z_j\},$$

then

$$E[T] \leq \left(\frac{8}{\delta^2}\right) \sum_{j=1}^{m-1} \left(\lambda \ln\left(\frac{6\delta\lambda}{4 + z_j \delta \lambda}\right) + \frac{1}{z_j}\right).$$

The level-based theorem gives an upper bound on the expected time until the algorithm discovers an element in the last level A_m, given that certain conditions on the operator D and population size λ are satisfied. Time is measured by the random variable T, which is defined as the number of individuals sampled in line 2 of Algorithm 1 until the end of the first generation having an individual in level A_m. This is an upper bound on the total number of sampled search points until a search point in A_m is conceived for the first time.

Informally, the two first conditions require a relationship between the current population P and the distribution $D(P)$ of the individuals in the next generation: Condition (G1) demands that the probability of creating an individual at level $j+1$ or higher is at least z_j when some fixed portion γ_0 of the population has reached level j or higher. Furthermore, if the number of individuals at level $j+1$ or higher is at least $\gamma\lambda > 0$, condition (G2) requires that their number tends to increase further, by a multiplicative factor of $1+\delta$. Finally, (G3) requires a sufficiently large population. When all conditions are satisfied, an upper bound on the expected time for the algorithm to create an individual in A_m is guaranteed.

We recommend these steps when applying the theorem:

1. Identify a partition of $\mathcal{X}$ reflecting the "typical" progress of the population to the target set A_m.
2. Find parameter settings of the algorithm and corresponding parameters γ_0 and δ which allow condition (G2) to be satisfied, possibly by adjusting the partition of $\mathcal{X}$.
3. For each level $j \in [m-1]$, estimate lower bounds z_j such that condition (G1) holds.
4. Determine the lower bound on the population size λ in (G3), using the parameters obtained in the previous steps.
5. Compute the upper bound on $E[T]$. The terms $\ln\left(\frac{6\delta\lambda}{4+z_j\delta\lambda}\right)$ can be bounded by underestimating the denominator, either by 4, or by $z_j\delta\lambda$, leading to the upper bounds $\ln(3\lambda/2)$, respectively $\ln(6/z_j)$.

We now illustrate this methodology on a simple example.

Corollary 2. *Algorithm 3 with* $\lambda \geq 72(\ln(m)+9)$ *produces less than* $216(m-1)(\lambda+1)$ *individuals in expectation before it discovers* m.

Algorithm 3 Example algorithm to illustrate Theorem 1.

1: Sample an initial population $P_0 \sim \mathrm{Unif}([m]^\lambda)$ u.a.r.
2: **for** $t = 0, 1, 2, \ldots$ until termination condition met **do**
3: Sort $P_t = (x_1, \ldots, x_\lambda)$ s.t. $x_1 \geq x_2 \geq \cdots \geq x_\lambda$.
4: **for** $i = 1$ to λ **do**
5: $z := x_k$, where $k \sim \mathrm{Unif}([\lambda/2])$.
6: $y := z + \mathrm{Unif}(\{-c, 0, 1\})$ for any fixed $c \in [m]$
7: $P_{t+1}(i) := \max\{1, \min\{y, m\}\}$.
8: **end for**
9: **end for**

To illustrate Theorem 1, we estimate the time until the element m is contained in the population of Algorithm 3. The search space $\mathcal{X}$ is the set of natural numbers between 1 and m. Following the scheme of Algorithm 1, the operator D corresponds to lines 3–6. The new individual y is obtained by first selecting uniformly at random one of the best $\lambda/2$ individuals in the population (lines 3 and 5) and mutating this individual by adding 1 subtracting c or doing nothing, with equal probabilities. We will see that the value of c does not matter in our analysis.

We now carry out the steps described previously.

Step 1: We use the partition $A_j := \{j\}$ for all $j \in [m]$.

Step 2: Assume that the *current level* is $j < m-1$. This means that in P_t, there are $\gamma_0\lambda$ individuals in $A_{\geq j}$, i. e., with fitness at least j, and at least $\gamma\lambda$ but less than $\gamma_0\lambda$ individuals in $A_{\geq j+1}$, i. e., with fitness at least $j+1$. We need to estimate $\Pr_{y\sim D(P_t)}(y \in A_{\geq j+1})$, i. e., the probability of producing an individual with fitness at least $j+1$. We say that a selection event is "good" if in line 5, the algorithm selects an individual in $A_{\geq j+1}$, i. e., with fitness at least $j+1$. If $\gamma \leq 1/2$, then the probability of a good selection event is at least $\gamma\lambda/(\lambda/2) = 2\gamma$. We say that a mutation event is "good" if in line 6, the algorithm does not subtract c from the selected search point. The probability of a good mutation event is $2/3$. Selection and mutation are independent events, hence we have shown for all $\gamma \in (0, 1/2]$ that $\Pr_{y\sim D(P_t)}(y \in A_{\geq j+1}) \geq (2\gamma)(2/3) = \gamma\left(1+\frac{1}{3}\right)$. Condition (G2) is therefore satisfied with $\delta = 1/3$ for any $\gamma_0 \leq 1/2$. We will choose the parameter γ_0 later.

Step 3: Assume that population P_t has at least $\gamma_0\lambda$ individuals in $A_{\geq j}$. The algorithm can then produce an individual in $A_{\geq j+1}$ by selecting an individual in $A_{\geq j}$, and mutate this individual into $A_{\geq j+1}$ by adding 1 in line 6. We can conveniently fix $\gamma_0 := 1/2$, so that the probability of selecting an individual in $A_{\geq j}$ becomes 1. Furthermore, the probability of adding 1 to the selected individual is exactly $1/3$. Hence, we have $\Pr_{y\sim D(P_t)}(y \in A_{\geq j+1}) \geq 1(1/3)$, and we can satisfy condition (G1) by defining $z_j := 1/3$ for all $j \in [m-1]$.

Step 4: For the parameters we have chosen, it is easy to see by numerical calculation that the population size $\lambda \geq 72(\ln(m)+9)$ satisfies condition (G3).

Step 5: Using that $\ln\left(\frac{6\delta\lambda}{4+\delta\lambda z_j}\right) < \ln\left(\frac{6}{z_j}\right)$, the expected time to discover the point m is no more than

$$\frac{8}{\left(\frac{1}{3}\right)^2}\sum_{j=1}^{m-1}\left(\lambda\ln\left(\frac{6}{\frac{1}{3}}\right)+\frac{1}{\frac{1}{3}}\right) < 216(m-1)(\lambda+1).$$

2.3 Proof of the level-based theorem

Theorem 1 will be proved using drift analysis [28, 29], which is a standard tool in theory of randomised search heuristics. We use the following variant of the additive drift theorem [29] and the proof can be found in the appendix. The lower bound statement will be used in the last part of paper to discuss the tightness of the level-based theorem. In what follows, "$(Z_{t+1} - Z_t + \varepsilon)\,; t < T_a$" is a short-hand notation for "$(Z_{t+1} - Z_t + \varepsilon) \cdot \mathbb{1}_{\{t<T_a\}}$" (see 6.3 in [57]).

Theorem 3 (Additive drift theorem)**.** *Let $(Z_t)_{t\in\mathbb{N}}$ be a discrete-time stochastic process in $[0,\infty)$ adapted to any filtration $(\mathscr{F}_t)_{t\in\mathbb{N}}$. Define $T_a := \min\{t \in \mathbb{N} \mid$*

$Z_t \leq a\}$ *for any* $a \geq 0$. *For some* $\varepsilon > 0$ *and constant* $0 < b < \infty$, *define the conditions*

1.1) $E[Z_{t+1} - Z_t + \varepsilon \,; t < T_a \mid \mathscr{F}_t] \leq 0$ *for all* $t \in \mathbb{N}$,

1.2) $E[Z_{t+1} - Z_t + \varepsilon \,; t < T_a \mid \mathscr{F}_t] \geq 0$ *for all* $t \in \mathbb{N}$,

2) $Z_t < b$ *for all* $t \in \mathbb{N}$, *and*

3) $E[T_a] < \infty$.

If 1.1), 2), and 3) hold, then $E[T_a \mid \mathscr{F}_0] \leq Z_0/\varepsilon$.
If 1.2), 2), and 3) hold, then $E[T_a \mid \mathscr{F}_0] \geq (Z_0 - a)/\varepsilon$.

When applying the additive drift theorem to a complex process, Z_t is the result of a (measurable) mapping of the states of the process to a real number. Such a mapping is called the *distance function*, which measures the distance to some target state. Our distance function takes into account both the current level of the population, as well as the distribution of the population around the current level. In particular, let the current level Y_t be the highest level $j \in [m]$ such that there are at least $\gamma_0\lambda$ individuals at level j or higher. Furthermore, for any level $j \in [m]$, let $X_t^{(j)}$ be the number of individuals at level j or higher. Hence, we describe the dynamics of the population by $m+1$ stochastic processes $X_t^{(1)}, \ldots, X_t^{(m)}, Y_t$. Assuming that these processes are adapted to a filtration $\mathscr{F}_t$, we write $E_t[X] := E[X \mid \mathscr{F}_t]$ and $\Pr_t(\mathcal{E}) := E[\mathbb{1}_{\mathcal{E}} \mid \mathscr{F}_t]$. Our approach is to measure the distance of the population at time t by a scalar $g(X_t^{(Y_t+1)}, Y_t)$, where g is a function that satisfies the conditions in Definition 4.

Definition 4. *A function* $g : (\{0\} \cup [\lambda]) \times [m] \to \mathbb{R}$ *is called a* level function *if the following three conditions hold*

1. $\forall x \in \{0\} \cup [\lambda], \forall y \in [m-1]: \quad g(x,y) \geq g(x,y+1)$,
2. $\forall x \in \{0\} \cup [\lambda-1], \forall y \in [m]: \quad g(x,y) \geq g(x+1,y)$,
3. $\forall y \in [m-1]: \quad g(\lambda,y) \geq g(0,y+1)$.

Note that the sum of two level functions is also a level function. Furthermore, the conditions ensure that the distance $g(X_t^{(Y_t+1)}, Y_t)$ of the population decreases monotonically with the current level Y_t. Lemma 5 shows that this monotonicity allows an upper bound on the distance in the next generation which is partly independent of the change in current level.

Lemma 5. *If* $Y_{t+1} \geq Y_t$, *then for any level function* g

$$g\left(X_{t+1}^{(Y_{t+1}+1)}, Y_{t+1}\right) \leq g\left(X_{t+1}^{(Y_t+1)}, Y_t\right).$$

Proof. The statement is trivial when $Y_t = Y_{t+1}$. On the other hand, if $Y_{t+1} > Y_t$, then the conditions in Definition 4 imply

$$\begin{aligned} g\left(X_{t+1}^{(Y_{t+1}+1)}, Y_{t+1}\right) &\leq g\left(0, Y_{t+1}\right) \leq g\left(0, Y_t + 1\right) \\ &\leq g\left(\lambda, Y_t\right) \leq g\left(X_{t+1}^{(Y_t+1)}, Y_t\right). \end{aligned}$$

□

Proof of Theorem 1. We will use Theorem 3 with respect to the parameter $a = 0$ and a stochastic process $Z_t := g\left(X_t^{(Y_t+1)}, Y_t\right)$, where g is a level-function to be defined, and $(Y_t)_{t\in\mathbb{N}}$ and $(X_t^{(j)})_{t\in\mathbb{N}}$ for $j \in [m]$ are stochastic processes to be defined. We consider the filtration $(\mathscr{F}_t)_{t\in\mathbb{N}}$ induced by the sequence of populations $(P_t)_{t\in\mathbb{N}}$.

We will assume w.l.o.g. that condition (G2) is also satisfied for $j = m-1$, for the following reason. Given Algorithm 1 with a certain mapping D, consider Algorithm 1 with a different mapping $D'(P)$: If $|P \cap A_m| = 0$, then $D'(P) = D(P)$; otherwise $D'(P)$ assigns probability mass 1 to some element x of P that is in A_m, e.g., to the first one among such elements. Note that D' meets conditions (G1) and (G2). Moreover, (G2) holds for $j = m-1$. For the sequence of populations $P'_0, P'_1, \ldots$ of Algorithm 1 with mapping D', we can put $T' := \lambda \cdot \min\{t \mid |P'_t \cap A_m| > 0\}$. Executions of the original algorithm and the modified one before generation T'/λ are identical. On generation T'/λ both algorithms place elements of A_m into the population for the first time. Thus, T' and T are equal in every realisation and their expectations are equal.

For any level $j \in [m]$ and time $t \geq 0$, let the random variable $X_t^{(j)} := |P_t \cap A_{\geq j}|$ denote the number of individuals in levels $A_{\geq j}$ at time t. Because $A_{\geq j}$ is partitioned into disjoint sets A_j and $A_{\geq j+1}$, the definition implies

$$|P_t \cap A_j| = X_t^{(j)} - X_t^{(j+1)} \tag{1}$$

Algorithm 1 samples all individuals in generation $t+1$ independently from distribution $D(P_t)$. Therefore, given the current population P_t, $X_{t+1}^{(j)}$ is binomially distributed $X_{t+1}^{(j)} \sim \mathrm{Bin}(\lambda, p_{t+1}^{(j)})$, where $p_{t+1}^{(j)} := \Pr_{t,y\sim D(P_t)}(y \in A_{\geq j})$ is the probability of sampling an individual in level j or higher.

The *current level* Y_t of the population at time t is defined as $Y_t := \max\left\{j \in [m] \;\middle|\; X_t^{(j)} \geq \gamma_0\lambda\right\}$. Note that $(X_t^{(j)})_{t\in\mathbb{N}}$ and $(Y_t)_{t\in\mathbb{N}}$ are adapted to the filtration $(\mathscr{F}_t)_{t\in\mathbb{N}}$ because they are defined in terms of the population process $(P_t)_{t\in\mathbb{N}}$.

When $Y_t < m$, there exists a unique $\gamma < \gamma_0$ such that

$$X_t^{(Y_t+1)} = |P_t \cap A_{\geq Y_t+1}| = \gamma\lambda, \tag{2}$$

$$X_t^{(Y_t)} = |P_t \cap A_{\geq Y_t}| \geq \gamma_0\lambda, \text{ and} \tag{3}$$

$$X_t^{(Y_t-1)} = |P_t \cap A_{\geq Y_t-1}| \geq \gamma_0\lambda. \tag{4}$$

In the case of $X_t^{(Y_t+1)} = 0$, it follows from (1), (2) and (3) that $|P \cap A_j| = X_t^{(Y_t)} \geq \gamma_0\lambda$. Condition (G1) for level $j = Y_t$ then gives

$$p_{t+1}^{(Y_t+1)} = \Pr_{y\sim D(P_t)}(y \in A_{\geq Y_t+1}) \geq z_{Y_t}. \tag{5}$$

Otherwise if $X_t^{(Y_t+1)} \geq 1$, conditions (G1) and (G2) for level $j = Y_t$ with (2) and (3) imply

$$p_{t+1}^{(Y_t+1)} = \Pr_{y\sim D(P_t)}(y \in A_{\geq Y_t+1}) \tag{6}$$

$$\geq \max\left\{(1+\delta)\frac{X_t^{(Y_t+1)}}{\lambda}, z_j\right\}. \tag{7}$$

Condition (G2) for level $j = Y_t - 1$ with (3) and (4) give

$$p_{t+1}^{(Y_t)} = \Pr_{y \sim D(P_t)}(y \in A_{\geq Y_t}) \geq (1+\delta)\gamma_0. \tag{8}$$

We now define the process $(Z_t)_{t\in\mathbb{N}}$ as $Z_t := 0$ if $Y_t = m$, and otherwise, if $Y_t < m$, we let $Z_t := g\left(X_t^{(Y_t+1)}, Y_t\right)$, where for all k, and for all $1 \leq j < m$, $g(k,j) = g_1(k,j) + g_2(k,j)$ and

$$\begin{aligned} g_1(k,j) &:= \ln\left(\frac{1+(\delta/2)\lambda}{1+(\delta/2)\max\{k, z_j\lambda/(1+\delta)\}}\right) \\ &\quad + \sum_{i=j+1}^{m-1} \ln\left(\frac{1+(\delta/2)\lambda}{1+(\delta/2)\lambda z_i/(1+\delta)}\right), \\ g_2(k,j) &:= \frac{1}{q_j}\left(1-\frac{\delta^2}{7}\right)^k + \sum_{i=j+1}^{m-1}\frac{1}{q_i}, \end{aligned}$$

where $q_j := 1-(1-z_j)^\lambda$.

It follows from Lemma 18 that $g(k,j)$ is a level function. Furthermore, $g(k,j) \geq 0$ for all $k \in \{0\} \cup [\lambda]$ and all $j \in [m]$. Due to properties 1 and 2 of level functions, and Lemma 31 from [14], the distance is always bounded from above by

$$\begin{aligned} g(0,1) &\leq \sum_{i=1}^{m-1}\left(\ln\left(\frac{1+(\delta/2)\lambda}{1+(\delta/2)z_i\lambda/(1+\delta)}\right) + \frac{1}{q_i}\right) \\ &< \sum_{i=1}^{m-1}\left(\ln\left(\frac{4+2\delta\lambda}{4+\delta z_i\lambda}\right) + 1 + \frac{1}{\lambda z_i}\right). \end{aligned} \tag{9}$$

Using that $z_i \leq 1$, this can be bounded further by

$$\begin{aligned} &< \sum_{i=1}^{m-1}\left(\ln\left(\frac{4+2\delta\lambda}{z_i(4+\delta\lambda)}\right) + 1 + \frac{1}{\lambda z_i}\right) \\ &= \sum_{i=1}^{m-1}\left(\ln\left(\frac{1}{z_i}\left(1+\frac{\delta\lambda}{4+\delta\lambda}\right)\right) + 1 + \frac{1}{\lambda z_i}\right). \end{aligned} \tag{10}$$

We then exploit that $\ln(x) \leq x-1$ for all $x > 0$ so

$$< \sum_{i=1}^{m-1}\left(\frac{2}{z_i} + \frac{1}{\lambda z_i}\right).$$

Finally, since $z_i \geq z_*$ and $\lambda \geq \lceil 4\ln(128)\rceil = 20$ by (G3)

$$< \frac{m}{z_*}\left(2+\frac{1}{\lambda}\right) \leq \frac{41m}{20z_*}. \tag{11}$$

Hence, we have $0 \leq Z_t < g(0,1) < \infty$ for all $t \in \mathbb{N}$ which implies that condition 2 of the drift theorem is satisfied.

The drift of the process at time t is $E_t[\Delta_{t+1}]$, where

$$\Delta_{t+1} := g\left(X_t^{(Y_t+1)}, Y_t\right) - g\left(X_{t+1}^{(Y_{t+1}+1)}, Y_{t+1}\right).$$

We bound the drift by the law of total probability as

$$\begin{aligned} E_t[\Delta_{t+1}] = (1 - \Pr_t(Y_{t+1} < Y_t)) E_t[\Delta_{t+1} \mid Y_{t+1} \geq Y_t] \\ + \Pr_t(Y_{t+1} < Y_t) E_t[\Delta_{t+1} \mid Y_{t+1} < Y_t]. \end{aligned} \tag{12}$$

The event $Y_{t+1} < Y_t$ holds if and only if $X_{t+1}^{(Y_t)} < \gamma_0\lambda$. Due to (8), we obtain the following by a Chernoff bound

$$\begin{aligned} \Pr_t(Y_{t+1} < Y_t) &= \Pr_t\left(X_{t+1}^{(Y_t)} < \left(1 - \frac{\delta}{1+\delta}\right)(1+\delta)\gamma_0\lambda\right) \\ &\leq \exp\left(-\frac{\delta^2\gamma_0\lambda}{2(1+\delta)}\right) \leq \frac{\delta^2 z_*}{128m}, \end{aligned} \tag{13}$$

where the last inequality takes into account the population size required by condition (G3). Given the low probability of the event $Y_{t+1} < Y_t$, it suffices to use the pessimistic bound (11)

$$E_t[\Delta_{t+1} \mid Y_{t+1} < Y_t] \geq -g(0,1) \geq -\frac{41m}{20z_*}. \tag{14}$$

If $Y_{t+1} \geq Y_t$, we can apply Lemma 5

$$\begin{aligned} &E_t[\Delta_{t+1} \mid Y_{t+1} \geq Y_t] \\ &\quad \geq E_t\left[g\left(X_t^{(Y_t+1)}, Y_t\right) - g\left(X_{t+1}^{(Y_t+1)}, Y_t\right) \mid Y_{t+1} \geq Y_t\right]. \end{aligned}$$

Note that event $Y_{t+1} \geq Y_t$ is equivalent to having $X_{t+1}^{(Y_t)} \geq \gamma_0\lambda$, then due to Lemma 20, in the following we can skip the condition on the event when needed.

If $X_t^{(Y_t+1)} = 0$, then $X_t^{(Y_t+1)} \leq X_{t+1}^{(Y_t+1)}$ and

$$E_t\left[g_1\left(X_t^{(Y_t+1)}, Y_t\right) - g_1\left(X_{t+1}^{(Y_t+1)}, Y_t\right) \mid Y_{t+1} \geq Y_t\right] \geq 0,$$

because the function g_1 satisfies property 2 in Definition 4. Furthermore, we have the lower bound

$$\begin{aligned} &E_t\left[g_2\left(X_t^{(Y_t+1)}, Y_t\right) - g_2\left(X_{t+1}^{(Y_t+1)}, Y_t\right) \mid Y_{t+1} \geq Y_t\right] \\ &\qquad > \Pr_t\left(X_{t+1}^{(Y_t+1)} \geq 1\right)(g_2(0, Y_t) - g_2(1, Y_t)) \geq \frac{\delta^2}{7}, \end{aligned}$$

where the last inequality follows because of (5) and $\Pr_t\left(X_{t+1}^{(Y_t+1)} \geq 1\right) \geq 1 - \left(1 - p_{t+1}^{(Y_t+1)}\right)^\lambda \geq 1 - (1 - z_{Y_t})^\lambda = q_{Y_t}$, and $g_2(0, Y_t) - g_2(1, Y_t) = (1/q_{Y_t})(\delta^2/7)$.

In the other case, where $X_t^{(Y_t+1)} \geq 1$, we obtain

$$E_t\left[g_1\left(X_t^{(Y_t+1)}, Y_t\right) - g_1\left(X_{t+1}^{(Y_t+1)}, Y_t\right) \mid Y_{t+1} \geq Y_t\right]$$

$$\geq E_t\left[\ln\left(\frac{1+\frac{\delta}{2}X_{t+1}^{(Y_t+1)}}{1+\frac{\delta}{2}\max\left\{X_t^{(Y_t+1)},\frac{z_{Y_t}\lambda}{1+\delta}\right\}}\right)\right] \geq \frac{\delta^2}{7},$$

where the last inequality follows from Lemma 19 for the parameters $X := X_{t+1}^{(Y_t+1)}$ and $p := p_{t+1}^{(Y_t+1)}$ as given by (7). For the function g_2, we get

$$E_t\left[g_2\left(X_t^{(Y_t+1)},Y_t\right)-g_2\left(X_{t+1}^{(Y_t+1)},Y_t\right)\mid Y_{t+1}\geq Y_t\right] = \frac{1}{q_{Y_t}}\left(\left(1-\frac{\delta^2}{7}\right)^{X_t^{(Y_t)}} - E_t\left[\left(1-\frac{\delta^2}{7}\right)^{X_{t+1}^{(Y_t+1)}}\right]\right) > 0,$$

where the last inequality is due to Lemma 6 from [14] (see also the supplementary materials), applied to $X_{t+1}^{(Y_t+1)} \sim \mathrm{Bin}(\lambda, p_{t+1}^{(Y_t+1)})$ with $p_{t+1}^{(Y_t+1)} \geq (1+\delta)X_t^{(Y_t+1)}/\lambda$ (see (7)) and the parameter $\kappa = -\ln(1-\delta^2/7) < \delta$.

Taking into account all cases, we have

$$E_t\left[\Delta_{t+1}\mid Y_{t+1}\geq Y_t\right] \geq \frac{\delta^2}{7}. \tag{15}$$

We now have bounds for all the quantities in (12) with (13), (14), and (15), and we get

$$\begin{aligned} E_t\left[\Delta_{t+1}\right] &= \left(1-\Pr_t\left(Y_{t+1}<Y_t\right)\right)E_t\left[\Delta_{t+1}\mid Y_{t+1}\geq Y_t\right] \\ &\quad + \Pr_t\left(Y_{t+1}<Y_t\right)E_t\left[\Delta_{t+1}\mid Y_{t+1}<Y_t\right] \\ &\geq \left(1-\frac{\delta^2 z_*}{128m}\right)\frac{\delta^2}{7} - \left(\frac{\delta^2 z_*}{128m}\right)\left(\frac{41m}{20z_*}\right) > \frac{\delta^2}{8}. \end{aligned}$$

We now verify condition 3 of Theorem 3, i. e., that T has finite expectation. Let $p_* := \min\{(1+\delta)/\lambda, z_*\}$, and note by conditions (G1) and (G2) that the current level increases by at least one with probability $\Pr_t\left(Y_{t+1}>Y_t\right) \geq (p_*)^{\gamma_0\lambda}$. Due to the definition of the modified process D', if $Y_t = m$, then $Y_{t+1} = m$. Hence, the probability of reaching $Y_t = m$ is lower bounded by the probability of the event that the current level increases in all of at most m consecutive generations, i. e., $\Pr_t\left(Y_{t+m}=m\right) \geq (p_*)^{\gamma_0\lambda m} > 0$. It follows that $E\left[T\right] < \infty$.

By Theorem 3 and the upper bound on $g(0,1)$ in (9),

$$E\left[T\right] \leq \lambda\cdot\frac{g(0,1)}{\delta^2/8} < \left(\frac{8\lambda}{\delta^2}\right)\sum_{i=1}^{m-1}\ln\left(\frac{4+2\delta\lambda}{4+z_i\delta\lambda}\right)+1+\frac{1}{\lambda z_i},$$

then using that $4 \leq \delta\lambda/5$ from (G3) and $(1/5+2)e < 6$,

$$< \left(\frac{8\lambda}{\delta^2}\right)\sum_{i=1}^{m-1}\left(\ln\left(\frac{6\delta\lambda}{4+z_i\delta\lambda}\right)+\frac{1}{\lambda z_i}\right). \qquad \square$$

It can be seen from the proof of Theorem 1 that it easily extends to algorithms where the mapping D is time-dependent, provided that (G1), (G2), and (G3) hold for any t for some fixed (time independent) values $z_1,\ldots,z_{m-1},\delta$, and γ_0.

3 Tools for Analysing Genetic Algorithms

In this section, we derive two corollaries of Theorem 1 tailored to Algorithm 2 and give conditions on tunable parameters of selection mechanisms making the corollaries applicable.

A fitness function is not defined explicitly in Algorithm 2, so no assumptions on an f-based partition will be needed in the corollaries. Here we generalise the *cumulative selection probability* of `select`, denoted $\beta(\gamma, P)$, which was defined relative to the fitness function f in [14], to the one that is relative to the partition $(A_1, \ldots, A_m)$. To define $\beta(\gamma, P)$ of `select` w.r.t. f for a population P of λ search points, we first assume $(f_1, \ldots, f_\lambda)$ to be the vector of sorted fitness values of P, i.e., $f_i \geq f_{i+1}$ for each $i \in [\lambda - 1]$. Then

$$\beta(\gamma, P) := \sum_{i=1}^{\lambda} p_{\text{sel}}(i \mid P) \cdot \left[f(P(i)) \geq f_{\lceil \gamma \lambda \rceil} \right]$$

for any $\gamma \in (0, 1]$. Here and below, $[\cdot]$ is the Iverson bracket.

Similarly, given a partition $(A_1, \ldots, A_m)$, if we use $(\ell_1, \ldots, \ell_\lambda)$ to denote the sorted levels of search points in P, i.e., $\ell_i \geq \ell_{i+1}$ for each $i \in [\lambda - 1]$, then the *cumulative selection probability* of `select` w.r.t. $(A_1, \ldots, A_m)$ is

$$\zeta(\gamma, P) := \sum_{i=1}^{\lambda} p_{\text{sel}}(i \mid P) \cdot \left[P(i) \in A_{\geq \ell_{\lceil \gamma \lambda \rceil}} \right].$$

These definitions are related by the following lemma, which is proved in the supplementary materials.

Lemma 6. *For any f-based partition of $\mathcal{X}$ and $\lambda \in \mathbb{N}$,*

$$\forall P \in \mathcal{X}^\lambda, \forall \gamma \in (0, 1] \quad \zeta(\gamma, P) \geq \beta(\gamma, P). \tag{16}$$

3.1 Analysis of non-permanent use of crossover

We first derive from Theorem 1 a corollary that is adapted to Algorithm 2 with $p_c < 1$. This setting covers the case $p_c = 0$, i.e., only unary variation operators are used. This specific case is the main subject of [14]. As we will see later on, stronger and more general results can be claimed with the corollary.

Corollary 7. *Given a partition $(A_1, \ldots, A_m)$ of $\mathcal{X}$, define $T := \min\{t\lambda \mid |P \cap A_m| > 0\}$ where for all $t \in \mathbb{N}$, $P_t \in \mathcal{X}^\lambda$ is the population of Algorithm 2 in generation t. If $p_c < 1$ and there exist $s_1, \ldots, s_{m-1}, s_*, p_0, \delta \in (0, 1]$, and a constant $\gamma_0 \in (0, 1)$ such that for any population $P \in \mathcal{X}^\lambda$,*

(M1) *for each level $j \in [m - 1]$*

$$p_{\text{mut}}\left(y \in A_{\geq j+1} \mid x \in A_j\right) \geq s_j$$

(M2) *for each level $j \in [m - 1]$*

$$p_{\text{mut}}\left(y \in A_{\geq j} \mid x \in A_j\right) \geq p_0$$

(M3) *for any population* $P \in (\mathcal{X} \setminus A_m)^\lambda$ *and* $\gamma \in (0, \gamma_0]$

$$\zeta(\gamma, P) \geq \frac{(1+\delta)\gamma}{p_0(1-p_c)}$$

(M4) *the population size* λ *satisfies*

$$\lambda \geq \left(\frac{4}{\gamma_0 \delta^2}\right) \ln\left(\frac{128m}{\gamma_0 s_* \delta^2}\right) \quad \textit{where } s_* := \min_{j \in [m-1]} \{s_j\},$$

then

$$E[T] < \left(\frac{8}{\delta^2}\right) \sum_{j=1}^{m-1} \left(\lambda \ln\left(\frac{6\delta\lambda}{4+\gamma_0 s_j \delta\lambda}\right) + \frac{1}{\gamma_0 s_j}\right).$$

Proof. We apply Theorem 1 following the guidelines. Step 1 is skipped because we already have the partition.

Step 2: Assume that $|P \cap A_{\geq j}| \geq \gamma_0 \lambda$ and $|P \cap A_{\geq j+1}| \geq \gamma\lambda > 0$ for some $\gamma \leq \gamma_0$. To create an individual in $A_{\geq j+1}$, it suffices to pick an $x \in |P \cap A_k|$ for any $k \geq j+1$ and mutate it to an individual in $A_{\geq k}$, the probability of such an event, according to (M2) and (M3), is at least $(1-p_c)\zeta(\gamma, P)p_0 \geq (1+\delta)\gamma$. So (G2) holds with p_0, δ and γ_0 from (M3).

Step 3: We are given $|P \cap A_j| \geq \gamma_0 \lambda$. Thus, with probability $\zeta(\gamma_0, P)$, the selection mechanism chooses an individual x in either A_j or $A_{\geq j+1}$. If $x \in A_j$, then the mutation operator will by (M1) upgrade x to $A_{\geq j+1}$ with probability s_j. If $x \in A_{\geq j+1}$, then by (M2), the mutation operator leaves the individual in $A_{\geq j+1}$ with probability p_0. Finally, no crossover occurs with probability $1 - p_c$, so the probability of producing an individual in $A_{\geq j+1}$ is at least $(1-p_c)\zeta(\gamma_0, P)\min\{s_j, p_0\} \geq (1-p_c)\zeta(\gamma_0, P)s_j p_0 > \gamma_0 s_j$ and (G1) holds with $z_j = \gamma_0 s_j$, $z_* = \gamma_0 s_*$.

Step 4: Given $z_* = \gamma_0 s_*$, condition (M4) yields (G3).

Step 5: Conditions (G1–3) are satisfied and Theorem 1 gives

$$\begin{aligned} E[T] &\leq \frac{8\lambda}{\delta^2} \sum_{j=1}^{m-1} \left(\ln\left(\frac{6\delta\lambda}{4+z_j\delta\lambda}\right) + \frac{1}{z_j\lambda}\right) \\ &= \frac{8}{\delta^2} \sum_{j=1}^{m-1} \left(\lambda \ln\left(\frac{6\delta\lambda}{4+\gamma_0 s_j \delta\lambda}\right) + \frac{1}{\gamma_0 s_j}\right). \end{aligned}$$

□

The proof implies that any operator can stand for crossover in line 4 of Algorithm 2, and the result will still hold.

3.2 Analysis of permanent use of crossover

We now adapt Theorem 1 to Algorithm 2 with $p_c = 1$.

Corollary 8. *Given a partition* $(A_1, \ldots, A_m)$ *of* $\mathcal{X}$*, define* $T := \min\{t\lambda \mid |P_t \cap A_m| > 0\}$*, where for all* $t \in \mathbb{N}$*,* $P_t \in \mathcal{X}^\lambda$ *is the population of Algorithm 2. If* $p_c = 1$ *and there exist* $s_1, \ldots, s_{m-1}, s_*, p_0, \varepsilon, \delta \in (0, 1]$*, and a constant* $\gamma_0 \in (0, 1)$ *such that*

(C1) *for each level* $j \in [m-1]$

$$p_{\text{mut}}\left(y \in A_{\geq j+1} \mid x \in A_j\right) \geq s_j,$$

(C2) *for each level* $j \in [m]$

$$p_{\text{mut}}\left(y \in A_{\geq j} \mid x \in A_j\right) \geq p_0,$$

(C3) *for each level* $j \in [m-2]$

$$p_{\text{xor}}\left(x \in A_{\geq j+1} \mid u \in A_{\geq j}, v \in A_{\geq j+1}\right) \geq \varepsilon,$$

(C4) *for any population* $P \in (\mathcal{X} \setminus A_m)^\lambda$ *and* $\gamma \in (0, \gamma_0]$

$$\zeta(\gamma, P) \geq \gamma \sqrt{\frac{1+\delta}{p_0 \gamma_0 \varepsilon}},$$

(C5) *the population size satisfies*

$$\lambda \geq \left(\frac{4}{\gamma_0 \delta^2}\right) \ln\left(\frac{128m}{\gamma_0 \delta^2 s_*}\right), \text{ where } s_* := \min_{j \in [m-1]} \{s_j\},$$

then

$$E\left[T\right] < \left(\frac{8}{\delta^2}\right) \sum_{j=1}^{m-1} \left(\lambda \ln\left(\frac{6\delta\lambda}{4 + \gamma_0 s_j \delta \lambda}\right) + \frac{1}{\gamma_0 s_j}\right).$$

Proof. We apply Theorem 1 following the guidelines. Again, Step 1 is skipped because the partition is already defined.

Step 2: We are given $|P \cap A_{\geq j}| \geq \gamma_0 \lambda$ and $|P \cap A_{\geq j+1}| \geq \gamma \lambda > 0$. To create an individual in $A_{\geq j+1}$, it suffices to pick the individual u in $A_{\geq j}$ and v in $A_{\geq j+1}$, then to produce an individual in A_k for any $k \geq j+1$ by crossover and not destroy the produced individual by mutation. The probability of such an event according to (C2), (C3), and (C4) is bounded from below by $\zeta(\gamma_0, P)\zeta(\gamma, P)\varepsilon p_0 \geq (1+\delta)\gamma$. Condition (G2) is then satisfied with the same γ_0 and δ as in (C4).

Step 3: We assume $|P \cap A_j| \geq \gamma_0 \lambda$. Note that condition (C3) written for level $j-1$ is $p_{\text{xor}}(x \in A_{\geq j} \mid u \in A_{\geq j-1}, v \in A_{\geq j}) \geq \varepsilon$, and because $A_{\geq j} \subset A_{\geq j-1}$ then $p_{\text{xor}}(x \in A_{\geq j} \mid u \in A_{\geq j}, v \in A_{\geq j}) \geq \varepsilon$. To create an individual in levels $A_{\geq j+1}$, it then suffices to pick both parents u and v from levels $A_{\geq j}$ in line 3, produce an intermediary offspring in A_k for any $k \geq j$ via crossover, and from this an individual in $A_{\geq j+1}$ via mutation. If $k = j$, we need to improve the produced individual by mutation, relying on (C1). Otherwise if $k > j$ it suffices not to destroy the produced individual by mutation, relying on (C2). It follows from (C4) that the probability of producing an individual in $A_{\geq j+1}$ is at least $\zeta(\gamma_0, P)^2 \varepsilon \min\{s_j, p_0\} \geq \zeta(\gamma_0, P)^2 \varepsilon s_j p_0 > \gamma_0 s_j$. Condition (G1) then holds for $z_j = \gamma_0 s_j$ and $z_* = \gamma_0 s_*$.

Step 4: Given that $z_* = \gamma_0 s_*$, (C5) implies (G3).

Step 5: Conditions (G1–3) are satisfied, so Theorem 1 gives

$$E\left[T\right] \leq \frac{8}{\delta^2} \sum_{j=1}^{m-1} \left(\lambda \ln\left(\frac{6\delta\lambda}{4 + \gamma_0 s_j \delta \lambda}\right) + \frac{1}{\gamma_0 s_j}\right). \qquad \square$$

Corollary 8 is similar to Corollary 7, except that condition (C2) has to additionally hold for level A_m, that (C3) is a new condition on the `crossover` operator, and that condition (C4) on the `select` operator differs from (M3).

3.3 Analysis of selection mechanisms

We show how to parametrise the following selection mechanisms such that condition (M3) of Corollary 7 and (C4) of Corollary 8 are satisfied. In *k-tournament selection*, k individuals are sampled uniformly at random with replacement from the population, and the fittest of these individuals is returned. In (μ, λ)-*selection*, parents are sampled uniformly at random among the fittest μ individuals in the population. A function $\alpha : \mathbb{R} \to \mathbb{R}$ is a ranking function [26] if $\alpha(x) \geq 0$ for all $x \in [0, 1]$, and $\int_0^1 \alpha(x)dx = 1$. In ranking selection with ranking function α, the probability of selecting an individual among $\gamma\lambda$ best individuals is $\int_0^\gamma \alpha(x)dx$. In *linear ranking selection*, parametrised by $\eta \in (1, 2]$, the ranking function is $\alpha(\gamma) := \eta(1 - 2\gamma) + 2\gamma$. We define *exponential ranking selection* parametrised by $\eta > 0$ with $\alpha(\gamma) := \eta e^{\eta(1-\gamma)}/(e^\eta - 1)$.

Lemma 9. *Assuming that $(A_1, \ldots, A_m)$ is a partition of $\mathcal{X}$ with $(A_1, \ldots, A_{m-1})$ being an f-based partition of $\mathcal{X} \setminus A_m$, for any constants $\delta' > 0$, $p_0 \in (0, 1)$, $\varepsilon \in (0, 1)$, and for any non-negative parameter $p_c = 1 - \Omega(1)$, there exists a constant $\gamma_0 \in (0, 1)$ such that all the following selection mechanisms*

1. *k-tournament selection,*
2. *(μ, λ)-selection,*
3. *linear ranking selection, and*
4. *exponential ranking selection*

with their parameters k, λ/μ, and η being set to no less than $\dfrac{1 + \delta'}{(1 - p_c)p_0}$ satisfy (M3), i. e., $\zeta(\gamma, P) \geq \dfrac{(1 + \delta'')\gamma}{p_0(1 - p_c)}$ for any $\gamma \in (0, \gamma_0]$, any $P \in (\mathcal{X}\setminus A_m)^\lambda$, and some constant $\delta'' > 0$.

Lemma 10. *Given a partition $(A_1, \ldots, A_m)$ of $\mathcal{X}$ with $(A_1, \ldots, A_{m-1})$ being an f-based partition of $\mathcal{X} \setminus A_m$, for any constants $\delta' > 0$, $p_0 \in (0, 1)$, and $\varepsilon \in (0, 1)$, there exists a constant $\gamma_0 \in (0, 1)$ such that the following selection mechanisms*

1. *k-tournament selection with $k \geq 4(1 + \delta')/(\varepsilon p_0)$,*
2. *(μ, λ)-selection with $\lambda/\mu \geq (1 + \delta')/(\varepsilon p_0)$, and*
3. *exponential ranking selection with $\eta \geq 4(1 + \delta')/(\varepsilon p_0)$*

satisfy (C4), i. e., $\zeta(\gamma, P) \geq \gamma\sqrt{\dfrac{1 + \delta'}{p_0\varepsilon\gamma_0}}$ for any $\gamma \in (0, \gamma_0]$, and any $P \in (\mathcal{X}\setminus A_m)^\lambda$.

Lemmas 9 and 10 are proved in the supplementary materials.

4 Applications to Genetic Algorithms

We now apply the results from Section 3. Proofs have been moved to the supplementary materials due to space limitations. Given a bitstring x a *bitwse mutation operator*, returns a bitstring y, where for each $i \in [n]$, bit y_i, is set independently to $1 - x_i$ with probability p_m and is otherwise kept equal to x_i. The parameter $p_m \in [0, 1]$ is called the *mutation rate*.

4.1 Optimisation of pseudo-Boolean functions

In this subsection, we consider the expected runtime of non-elitist GAs in Algorithm 2 on the following functions: $\textsc{OneMax}(x) := \sum_{i=1}^{n} x_i = |x|_1 = \textsc{Om}(x)$, $\textsc{LeadingOnes}(x) := \sum_{i=1}^{n} \prod_{j=1}^{i} x_i = \textsc{Lo}(x)$,

$$\textsc{Jump}_r(x) := \begin{cases} n+1 & \text{if } |x|_1 = n \\ r + |x|_1 & \text{if } |x|_1 \leq n-r \\ n - |x|_1 & \text{otherwise} \end{cases},$$

$\textsc{RoyalRoad}_r(x) := \sum_{i=0}^{n/r-1} \prod_{j=1}^{r} x_{ir+j}$, $\textsc{Linear}(x) := \sum_{i=1}^{n} c_i x_i$, where each $c_i \in \mathbb{R}$. For Linear, w.l.o.g. we can assume $c_1 \geq \cdots \geq c_n > 0$ [14]. We also consider the class of ℓ-Unimodal functions, where each function has exactly ℓ distinctive fitness values $f_1 < \cdots < f_\ell$, and each bitstring x of the search space is either optimal or it has a Hamming-neighbour y with $f(y) > f(x)$.

Several results about the runtime of EAs with parent or offspring population size greater than can be found in the literature. For the illustrative purpose, we cite just some of results. In [48], it is shown that $(1, \lambda)$ EA on OneMax has the runtime $\mathcal{O}(n \log n + n\lambda)$, provided that $\lambda \geq \log_{\frac{e}{e-1}} n$, and on the ℓ-Unimodal functions this algorithm has the runtime $O(\ell n + \ell\lambda)$, given that $\lambda \geq \log_{\frac{e}{e-1}}(\ell n)$. A non-elitist GA using bitwise mutation and tournament selection with $k = \Omega(\lambda)$ and $\lambda = \Omega(n \log n)$ has runtime $\mathcal{O}(n\lambda)$ on LeadingOnes [22]. The $(1 + \lambda)$ EA on any linear function has runtime $O(n \log n + n\lambda)$, see [18]. A $(\mu + 1)$ GA where the uniform crossover is applied with probability $p_c = \mathcal{O}(1/(nr))$ and μ is chosen appropriately is shown to have $\mathcal{O}\left(\mu n^2 r^3 + 4^r/p_c\right)$ runtime on $\textsc{Jump}_r$ function [31]. The case of $p_c = 1$ is treated in [10].

Our results presented below apply only to non-elitist GAs with bitwise mutation. For a moderate use of crossover, i.e., $p_c = 1 - \Omega(1)$, Corollary 7 and Lemma 9 yield

Theorem 11. *The expected runtime of the GA in Algorithm 2, with $p_c = 1 - \Omega(1)$, using any crossover operator, a bitwise mutation with mutation rate χ/n for any fixed constant $\chi > 0$ and one of the selection mechanisms: k-tournament selection, (μ, λ)-selection, linear or exponential ranking selection, with their parameters k, λ/μ, and η being set to no less than $(1+\delta)e^{\chi}/(1-p_c)$, where $\delta \in (0, 1]$ being any constant, is*

- $\mathcal{O}(n\lambda)$ *on* OneMax *if* $\lambda \geq c \ln n$,
- $\mathcal{O}\left(n^2 + n\lambda \ln \lambda\right)$ *on* LeadingOnes *if* $\lambda \geq c \ln n$,
- $\mathcal{O}(n\ell + \ell\lambda \ln \lambda)$ *on* ℓ-Unimodal *if* $\lambda \geq c \ln(\ell n)$,
- $\mathcal{O}\left(n^2 + n\lambda \ln \lambda\right)$ *on* Linear *if* $\lambda \geq c \ln n$,
- $\mathcal{O}\left(\left(\frac{n}{\chi}\right)^r + n\lambda + \lambda \ln \lambda\right)$ *on* $\textsc{Jump}_r$ *if* $\lambda \geq cr \ln n$,
- $\mathcal{O}\left(\left(\frac{n}{\chi}\right)^r \ln\left(\frac{n}{r}\right) + \frac{n\lambda \ln \lambda}{r}\right)$ *on* $\textsc{RoyalRoad}_{r \geq 2}$ *if* $\lambda \geq cr \ln n$,

for some sufficiently large constant c.

The proof is in the supplementary materials.

In the case of regular use of crossover, i. e., $p_c = 1$, we limit our consideration to *mask-based* crossovers. Given two parent genotypes u and v, such operator consists in first choosing (deterministically or randomly) a binary string $\tilde{m} = (m_1, \ldots, m_n)$ to produce two offspring vectors x', x'' as

$$x'_i = \begin{cases} u_i, & \text{if } m_i = 1 \\ v_i, & \text{otherwise,} \end{cases} \qquad x''_i = \begin{cases} v_i, & \text{if } m_i = 1 \\ u_i, & \text{otherwise.} \end{cases}$$

Then one element of $\{x', x''\}$ chosen uniformly at random is returned. The well-known uniform crossover and k-point crossover are examples of mask-based crossover operators.

For a frequent use of crossover, i. e., $p_c = 1$, Lemma 2 from [9], Corollary 8, and Lemma 10 yield

Theorem 12. *Assume that the GA in Algorithm 2 with $p_c = 1$ uses any mask-based crossover operator, a bitwise mutation with mutation rate χ/n for any fixed constant $\chi > 0$, and one of the following selection mechanisms:*

- *k-tournament selection with $k \geq 8(1+\delta)e^{\chi}$,*
- *(μ, λ)-selection with $\lambda/\mu \geq 2(1+\delta)e^{\chi}$, or*
- *exponential ranking selection with $\eta \geq 8(1+\delta)e^{\chi}$,*

for any constant $\delta > 0$. Then there exists a constant $c > 0$, such that the expected runtime of the GA is

- *$\mathcal{O}(n\lambda)$ on* OneMax *if $\lambda \geq c \ln n$,*
- *$\mathcal{O}\left(n^2 + n\lambda \ln \lambda\right)$ on* LeadingOnes *if $\lambda \geq c \ln n$.*

The proof can be found in the supplementary materials.

In the next sections, we further demonstrate the generality of Theorem 1 through Corollary 7 by deriving bounds on the expected runtime of GAs with $p_c = 1 - \Omega(1)$ to optimise or to approximate the optimal solutions.

4.2 Optimisation on permutation space

Given n distinct elements from a totally ordered set, we consider the problem of ordering them so that some *measure of sortedness* is maximised. In [49], the (1+1) EA was analysed on several measures of sortedness, including $\textsc{Inv}(\pi)$ which is defined to be the number of pairs (i, j) such that $1 \leq i < j \leq n$, $\pi(i) < \pi(j)$ (i. e., pairs in correct order). We show that with the method introduced in this paper, analysing GAs on Sorting problem with Inv measure, denoted by $\textsc{Sorting}_{\textsc{Inv}}$, is not much harder than analysing the (1+1) EA.

For the mutation, we use the Exchange(π) operator [49], which consecutively applies N pairwise exchanges between uniformly selected pairs of indices, where N is a random number drawn from a Poisson distribution with parameter 1.

Theorem 13. *If the GA in Algorithm 2 with $p_c = 1 - \Omega(1)$ uses any crossover operator, the* Exchange *mutation operator, one of the selection mechanisms k-tournament selection, (μ, λ)-selection, and linear or exponential ranking selection, with their parameters k, λ/μ and η being set to no less than $(1+\delta)e/(1-p_c)$, then there exists a constant $c > 0$ such that if the population size is $\lambda \geq c \ln n$, the expected time to obtain the optimum of $\textsc{Sorting}_{\textsc{Inv}}$ is $\mathcal{O}\left(n^2\lambda\right)$.*

The proof may be found in the supplementary materials.

4.3 Search for Local Optima

A great interest in the area of combinatorial optimisation is to find approximate solutions to NP-hard problems, because exact solutions for such problems are unlikely be computable in polynomial time under the so-called P≠NP hypothesis. In the case of maximisation problems, a feasible solution is called a *ρ-approximate solution* if its objective function value is at least ρ times the optimum for some $\rho \in (0, 1]$. Local search is one method among others to approximate solutions for combinatorial optimisation problems through finding local optima (a formal definition is given below). For a number of well-known problems, it was shown [1] that any local optimum is guaranteed to be a ρ-approximate solution with a constant ρ.

Suppose that a *neighbourhood* $\mathcal{N}(x) \subseteq \mathcal{X}$ is defined for every $x \in \mathcal{X}$. The mapping $\mathcal{N} : \mathcal{X} \to 2^{\mathcal{X}}$ is called the *neighbourhood mapping* and all elements of $\mathcal{N}(x)$ are called *neighbours* of x. For example, a frequently used neighbourhood mapping in the case of binary search space $\mathcal{X} = \{0, 1\}^n$ is defined by the Hamming distance $H(\cdot, \cdot)$ and a radius r as $\mathcal{N}(x) = \{y \mid H(x, y) \leq r\}$. If $f(y) \leq f(x)$ holds for all neighbours y of $x \in \mathcal{X}$, then x is called a *local optimum* w.r.t. $\mathcal{N}$. The set of all local optima is denoted by $\mathcal{LO}$ (note that global optima also belong to $\mathcal{LO}$).

A local search method starts from some initial solution y_0. Each iteration of the algorithm consists of moving from the current solution x to a new solution in its neighbourhood, so that the value of the fitness function is increased. The algorithm continues until a local optimum is reached. Let m be the number of different fitness values attained by solutions from $\mathcal{X} \backslash \mathcal{LO}$ plus 1. Then starting from any point, the local search method finds a local optimum within at most $m - 1$ steps, each step requiring at most $\mathcal{N}(x)$ fitness evaluations.

The following result provides sufficient conditions that ensure the GA finds (at least) a local optimum with a runtime not much greater than that of the local search.

Corollary 14. *Given some positive constants p_0, ε_0 and δ, define the following conditions:*

(X1) $p_{\mathrm{mut}}(y \mid x) \geq s$ *for any* $x \in \mathcal{X},\ y \in \mathcal{N}(x)$.

(X2) $p_{\mathrm{mut}}(x \mid x) \geq p_0$ *for all* $x \in \mathcal{X}$.

(X3) $p_{\mathrm{xor}}\big(f(x') \geq \max\{f(u), f(v)\} \mid u, v\big) \geq \varepsilon_0$ *for any* $u, v \in \mathcal{X}$.

(X4.1) *The non-elitist GA in Algorithm 2 is set with $p_{\mathrm{c}} = 1$, and it uses one of the following selection mechanisms:*

- *k-tournament selection with $k \geq \frac{4(1+\delta)}{\varepsilon_0 p_0}$,*
- *(μ, λ)-selection with $\frac{\lambda}{\mu} \geq \frac{(1+\delta)}{\varepsilon_0 p_0}$,*
- *exponential ranking selection with $\eta \geq \frac{4(1+\delta)}{\varepsilon_0 p_0}$.*

(X4.2) *The non-elitist GA is set with $p_{\mathrm{c}} = 1 - \Omega(1)$, and it uses one of the following selection mechanisms: k-tournament selection, (μ, λ)-selection,*

linear or exponential ranking selection, with their parameters k, λ/μ, and η being set to no less than $\frac{(1+\delta)}{(1-p_c)p_0}$.

If (X1–3) and (X4.1) hold, or exclusively (X1–2) and (X4.2) hold, then there exists a constant c, such that for $\lambda \geq c \ln\left(\frac{m}{s}\right)$, a local optimum is reached for the first time after $\mathcal{O}\left(m\lambda \ln \lambda + \frac{m}{s}\right)$ fitness evaluations in expectation.

Condition (X4.1) or (X4.2) characterises the setting of selection mechanisms, while (X1–3) bear the properties of the variation operators over the neighbourhood structure $\mathcal{N}$. Particularly, (X1) assumes a lower bound s on the probability that the mutation operator transforms an input solution into a specific neighbour. Note that in most of the local search algorithms, the neighbourhood $\mathcal{N}(x)$ may be enumerated in polynomial time of the problem input size. For such neighbourhood mappings, a mutation operator that generates the uniform distribution over $\mathcal{N}(x)$ will satisfy (X1) with $1/s$ polynomially bounded in the problem input size.

If crossover is frequently used, i. e., $p_c = 1$, we also need to satisfy condition (X3) on the the crossover operator. It requires that the fitness of solution x on the output of crossover is not less than the fitness of parents with probability at least ε_0. Note that such a requirement is satisfied with $\varepsilon_0 = 1$ for the optimized crossover operators, where the offspring is computed as a solution to the *optimal recombination problem* (see, e.g., [23]). This supplementary problem is known to be polynomially solvable for Maximum Clique, Set Packing, Set Partition and some other NP-hard problems (see e.g. [23]).

The proof of Corollary 14 directly follows from Corollaries 8 and 7 of Theorem 1, see supplementary materials.

Consider the binary search space $\{0,1\}^n$ with Hamming neighbourhood of a constant radius r, a fitness function f such that $m \in \text{poly}(n)$, and assume that the GA uses the bitwise mutation operator and $p_c = 1 - \Omega(1)$. This operator outputs a string y, given a string x, with probability $p_m^{H(x,y)}(1-p_m)^{n-H(x,y)}$. Note that probability $p_m^j(1-p_m)^{n-j}$, as a function of p_m attains its maximum at $p_m = j/n$. It is easy to show (see, e. g., [22]) that for any $x \in \mathcal{X}$ and $y \in \mathcal{N}(x)$, the bitwise mutation operator with $p_m = r/n$ satisfies the condition $p_{\text{mut}}(y \mid x) = \mathcal{O}(1/n^r)$. For a sufficiently large n and any $x \in \mathcal{X}$ holds $p_{\text{mut}}(x \mid x) \geq e^{-r}/2 = \Omega(1)$. By Corollary 14, a GA with the above mentioned operators, given appropriate λ, p_m and p_c, first visits a local optimum w. r. t. a constant radius Hamming neighbourhood after a polynomially bounded number of fitness evaluations in expectation.

Consider the following two unconstrained (and unweighted) problems. (i) Max-SAT: given a CNF formula in n logical variables which is represented by m' clauses $\mathbf{c}_1, \ldots, \mathbf{c}_{m'}$ and each clause is a disjunction of logical variables or their negations, it is required to find an assignment of the variables so that the number of satisfied clauses is maximised. (ii) Max-CUT: given an undirected graph $G = (V, E)$, it is required to find a partition of V into two sets $(S, V \setminus S)$, so that $\delta(S) := |\{(u,v) \mid (u,v) \in E, u \in S, v \notin S\}|$, is maximised.

Both problems are NP-hard, and their solutions can be naturally represented by bitstrings. Particularly, any local optimum w. r. t. the neighbourhood defined by Hamming distance 1 has at least half the optimal fitness [1].

Theorem 15. *Suppose the GA in Algorithm 2 is applied to* Max-SAT *or to* Max-CUT *using a bitwise mutation with $p_m = \chi/n$, where $\chi > 0$ is a con-*

stant, a crossover with $p_c = 1 - \Omega(1)$ *and one of the selection mechanisms: k-tournament selection,* (μ, λ)*-selection, linear or exponential ranking selection, with their parameters* k*,* λ/μ *and* η *being set to no less than* $\frac{(1+\delta)e^{\chi}}{1-p_c}$*, where* $\delta > 0$ *is any constant. Then there exists a constant* c*, such that for* $\lambda \geq c \ln(nm')$*, a* $1/2$*-approximate solution is reached for the first time after* $\mathcal{O}\left(m'\lambda \ln \lambda + nm'\right)$ *fitness evaluations in expectation for* Max-SAT*, and after* $\mathcal{O}\left(|E|\lambda \ln \lambda + |V|\,|E|\right)$ *for* Max-Cut*.*

The proof is analogous to the analysis of ℓ-Unimodal function in Theorem 11, combined with Corollary 14, where $m \leq m' + 1$ for Max-SAT and $m \leq |E| + 1$ for Max-Cut.

5 Estimation of Distribution Algorithms

There are few rigorous runtime results for UMDA and other EDAs. The techniques used in previous analyses of EDAs were often complex, e. g., relying on Markov chains theory. Surprisingly, even apparently simple problems, such as the expected runtime of UMDA on OneMax, were open until recently. Indeed, much more is known about classical EAs, e. g., the (1+1) EA solves OneMax in expected time $\Theta(n \ln n)$, and this is optimal for the class of unary unbiased black-box algorithms [38].

Algorithm 1 matches closely the typical behaviour of estimation of distribution algorithms: given a current distribution over the search space, sample a finite number of search points, and update the probability distribution. We demonstrate the ease at which the expected runtime of UMDA with margins and truncation selection on the OneMax function can be obtained using the level-based theorem without making any simplifying assumptions about the optimisation process.

To start, we give a formal description of UMDA. If $P \in \mathcal{X}^{\lambda}$ is a population of λ solutions, let $P(k, i)$ denote the value in the i-th bit position of the k-th solution in P. The Univariate Marginal Distribution Algorithm (UMDA) with (μ, λ)-truncation selection is defined in Algorithm 4.

Algorithm 4 UMDA

Require:
A pseudo-Boolean function $f : \{0,1\}^n \to \mathbb{R}$,
and "margins" $m' \in [0, \mu/2)$.

1: Initialise the vector $p_0 := (1/2, \ldots, 1/2)$.
2: **for** $t = 1, 2, 3, \ldots$ **do**
3: **for** $x = 1$ to λ **do**
4: Sample the x-th individual $P_t(x, \cdot)$ according to

$$P_t(x, i) \sim \text{Bernoulli}(p_{t-1}(i)) \text{ for all } i \in [n].$$

5: **end for**
6: Sort the population P_t according to f.
7: For all $i \in [n]$, with $X_i := \sum_{k=1}^{\mu} P_t(k, i)$, define

$$p_t(i) := \max\left\{\min\left\{\frac{X_i}{\mu}, 1 - \frac{m'}{\mu}\right\}, \frac{m'}{\mu}\right\}.$$

8: **end for**

The algorithm has three parameters, the parent population size μ, the offspring population size λ, and a parameter $m' < \mu$ controlling the size of the margins. It is necessary to set $m' > 0$ to prevent a premature convergence, e. g., without this margin, $p_t(i)$ can go to a non-optimal fixation, this prevents further exploration and causes an infinite runtime. Based on insights about optimal mutation rates in the (1+1) EA, we will use the parameter setting $m' = \mu/n$ in the rest of this section.

It is immediately clear that the UMDA in Algorithm 4 is a special case of Algorithm 1. The probability distribution $D(P_t)$ of y is computed in steps 7, and is $\Pr(y = x) = \prod_{j=1}^{n} p_t(j)^{x_j} (1 - p_t(j))^{1-x_j}$. for any bitstring $x \in \{0,1\}^n$.

In some other randomised search heuristics, such as ant colony optimisation (ACO) and compact genetic algorithms (cGA), the sampling distribution D_t does not only depend on the current population, but also on additional information, such as pheromone values. The level-based theorem does not apply to such algorithms.

Theorem 16. *Given any positive constants $\delta \in (0,1)$, and $\gamma_0 \leq \frac{1}{(1+\delta)13e}$, the UMDA with offspring population size λ with $b\ln(n) \leq \lambda \leq n/\gamma_0$ for some constant $b > 0$, parent population size $\mu = \gamma_0\lambda$, and margins $m' = \mu/n$, has expected optimisation time $\mathcal{O}(n\lambda \ln \lambda)$ on* OneMax.

The proof which is available in our supplementary materials follows our guidelines of level-based analysis, a preliminary version of it appeared in [13]. To obtain lower bounds on the tail of the level distribution, we make use of the Feige inequality [24] (or see Corollary 3 in [13]).

A similar analysis for LeadingOnes [13] yields a runtime bound $\mathcal{O}\left(n\lambda \ln \lambda + n^2\right)$ with offspring population size $\lambda \geq b\ln(n)$ for some constant $b > 0$ without use of Feige's inequality. The previous result [8] on LeadingOnes requires a larger population size and gives a longer runtime bound.

Table 1 summarises the runtime bounds for the example applications of the tools presented in this paper and the above mentioned result for UMDA on

LeadingOnes.

Table 1: Summary of results for the GA (Algo. 2, $p_c = 1 - \Omega(1)$), GA1 (Algo. 2, $p_c = 1$) and UMDA (Algo. 4, $m' = \mu/n$).

Runtime result

Problem	**Algorithm**	**Min. λ**	**Runtime**
OneMax	GA, GA1	$c \ln n$	$\mathcal{O}(n\lambda)$
	UMDA	$c \ln n$	$\mathcal{O}(n\lambda \ln \lambda)$
LeadingOnes	GA, GA1, UMDA	$c \ln n$	$\mathcal{O}\left(n^2 + n\lambda \ln \lambda\right)$
ℓ-Unimodal	GA	$c \ln(n\ell)$	$\mathcal{O}(n\ell + \ell\lambda \ln \lambda)$
Linear	GA	$c \ln n$	$\mathcal{O}\left(n^2 + n\lambda \ln \lambda\right)$
Jump$_r$	GA	$cr \ln n$	$\mathcal{O}\left(\left(\frac{n}{\chi}\right)^r + n\lambda + \lambda \ln \lambda\right)$
RoyalRoad$_{r \geq 2}$	GA	$cr \ln n$	$\mathcal{O}\left(\left(\frac{n}{\chi}\right)^r \ln\left(\frac{n}{r}\right) + \frac{n\lambda \ln \lambda}{r}\right)$
Sorting$_{\text{Inv}}$	GA	$c \ln n$	$\mathcal{O}\left(n^2\lambda\right)$
$\frac{1}{2}$-approx. Max-SAT	GA	$c \ln(nm')$	$\mathcal{O}(nm' + m'\lambda \ln \lambda)$
$\frac{1}{2}$-approx. Max-CUT	GA	$c \ln(\lvert V\rvert \, \lvert E\rvert)$	$\mathcal{O}(\lvert V\rvert \, \lvert E\rvert + \lvert E\rvert\lambda \ln \lambda)$

Configuration

Alg.	**Recomb.**	**Selection**	**Setting**
GA	any	k-tournament	$k \geq \frac{(1+\delta)e^\chi}{1-p_c}$
	any	(μ, λ)-selection	$\frac{\lambda}{\mu} \geq \frac{(1+\delta)e^\chi}{1-p_c}$
	any	linear ranking	$\eta \geq \frac{(1+\delta)e^\chi}{1-p_c}$
	any	exp. ranking	$\eta \geq \frac{(1+\delta)e^\chi}{1-p_c}$
GA1	mask-based	k-tournament	$k \geq 8(1+\delta)e^\chi$
	mask-based	(μ, λ)-selection	$\frac{\lambda}{\mu} \geq 2(1+\delta)e^\chi$
	mask-based	exp. ranking	$\eta \geq 8(1+\delta)e^\chi$
UMDA	n/a	(μ, λ)-selection	$\frac{\lambda}{\mu} \geq 13(1+\delta)e$

On $\{0,1\}^n$, GA and GA1 use bitwise mutation operator with rate χ/n, where χ is any constant. On permutation search space, i.e., Sorting, GA uses Exchange mutation and its setting assumes $\chi = 1$. In the case of Max-SAT, n is the number of logical variables and m' is the number of clauses. Parameter δ is any positive constant, and c is some constant.

6 The level-based theorem is almost tight

How accurate are the time bounds provided by the level-based theorem? To answer this question, we first interpret the theorem as a universally quantified statement over the operators D satisfying the conditions of the theorem. More formally, given a choice of level-partition and set of parameters $z_1, \ldots, z_{m-1}, \delta, \gamma_0$, which we collectively denote by Θ, the theorem can be expressed in the form of $\forall D \in \mathcal{D}_\Theta \colon E[T_D] \leq t_\Theta$, where $\mathcal{D}_\Theta$ is the set of operators D in Algorithm 1 that satisfy the conditions of the level-based theorem with parameterisation Θ, $E[T_D]$ is the expected runtime of Algorithm 1 with a given operator D, and t_Θ is the upper time bound provided by the level-based theorem which depends on Θ.

In order to obtain an accurate bound for a specific operator D, e. g., the (μ,λ) EA applied to the OneMax function, it is necessary to choose a parametrisation Θ that reflects this process as tightly as possible. If the bounds on the upgrade probabilities z_j for the (μ,λ) EA are too small, then the class $\mathcal{D}_\Theta$ includes other processes which are slower than the (μ,λ) EA, and the corresponding bound t_Θ cannot be accurate. Hence, the theorem is limited by the accuracy at which one can describe the process by some class $\mathcal{D}_\Theta$.

Assuming a fixed parameterisation Θ, it is possible to make a precise statement about the tightness of the upper bound t_Θ. Theorem 17 below is an existential statement on the form $\exists D \in \mathcal{D}_\Theta \colon E[T_D] \geq t'_\Theta$,where the lower bound t'_Θ is close to the upper bound t_Θ. Hence, given the information the theorem has about the process through Θ, the runtime bound is close to optimal. More information about the process would be required to obtain a more accurate runtime bound.

In some concrete cases, one can prove that the level-based theorem is close to optimal, using parallel black-box complexity theory [2]. From Corollary 7 with $p_c = 0$, which specialises the level-based theorem to algorithms with unary mutation operators, one can obtain the bounds $\mathcal{O}(n\lambda + n \ln n)$ for OneMax, and $\mathcal{O}\left(n\lambda \ln \lambda + n^2\right)$ for LeadingOnes for appropriately parameterised EAs. These bounds are within a $\mathcal{O}(\ln \lambda)$-factor of the lower bounds that hold for any parallel unbiased black-box algorithm [2]. For population sizes $\lambda = \mathcal{O}(n/\ln n)$ and $\lambda = \Omega(\ln n)$, the resulting $\mathcal{O}\left(n^2\right)$ bound on LeadingOnes is asymptotically tight, because it matches the lower bound that holds for all black-box algorithms with unary unbiased variation operators [38].

Theorem 17. *Given any partition of $\mathcal{X}$ into m non-empty subsets $(A_1, \ldots, A_m)$, for any $z_1, \ldots, z_{m-1}, \delta, \gamma_0 \in (0,1)$, where $z_j \in (0, \gamma_0)$ for all $j \in [m-1]$, and $\lambda \in \mathbb{N}$, there exists a mapping D which satisfies conditions (G1), (G2), and (G3) of Theorem 1, such that Algorithm 1 with mapping D has expected hitting time*

$$E[T] \geq \left(\frac{2}{3\delta} \sum_{j=1}^{m-2} \lambda \ln \left(\frac{\gamma_0 \lambda}{1 + 2\lambda z_j + 1/\delta^2} \right) \right) + \sum_{j=1}^{m-1} \frac{1}{z_j},$$

where $T := \min\{\lambda t \in \mathbb{N} \mid |P_t \cap A_m| > 0\}$.

The proof can be found in the supplementary materials.

7 Conclusion

This paper introduces a new technique, the so-called level-based analysis, that easily yields upper bounds on the expected runtime of complex, non-elitist search processes. The technique was first illustrated on Genetic Algorithms. We have shown that GAs efficiently optimise standard benchmark functions and some combinatorial optimisation problems. As long as the population size is not overly large, the population does not incur an asymptotic slowdown on these functions compared to standard EAs that do not use populations. So, speedups can be achieved by parallellising fitness evaluations. Furthermore, previous work using a weaker form of the level-based analysis indicates that non-elitist, population-based EAs have an advantage on more complex problems, including those with noisy [12], dynamic [11], and peaked [15] fitness landscapes. To demonstrate the generality of the theorem, we also provided runtime results for the UMDA algorithm, an Estimation of Distribution Algorithm, for which few theoretical results exist. Finally, we have shown via lower bounds on the runtime of a concrete process that, given the information the theorem requires about the process, the upper bounds are close to tight.

The conditions of the level-based theorem yield settings for algorithmic parameters, such as population size, mutation and crossover rates, selection pressure etc., that are sufficient to guarantee a time complexity bound. This opens up the possibility of theory-led design of EAs with guaranteed runtimes, where the algorithm is designed to satisfy the conditions of the level-based theorem. This paper also opens several new directions for future work. An important open problem is to develop techniques for proving lower bounds on the runtime of Algorithm 1. Rowe and Sudholt showed that the non-elitist $(1,\lambda)$ EA becomes inefficient when the population size is too small [48]. While condition (G3) in this paper gives a sufficient condition for the population size, it would be interesting to determine a necessary condition on the population size to efficiently reach the last level A_m.

Acknowledgements The research was supported by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 618091 (SAGE) and Russian Foundation for Basic Research grants 15-01-00785 and 16-01-00740. Early ideas were discussed at Dagstuhl Seminars 13271 and 15211 "Theory of Evolutionary Algorithms".

Appendix A

We prove the additive drift theorem. Equalities and inequalities involving conditional expectation w. r. t. a σ-algebra, (e. g., $E[X \mid \mathscr{F}] \leq Y$), hold almost surely. The proof relies on martingale theory (see, e. g., [57]).

Proof of Theorem 3. We first show the upper bound statement. Define the stopped process $S_t := Z_{t \wedge T_a} + \varepsilon(t \wedge T_a)$, where $t \wedge T_a := \min(t, T_a)$. By the definition of this process, it holds for all $t \in \mathbb{N}$ almost surely that

$$|S_t| \leq b + \varepsilon T_a, \tag{17}$$

and, hence by condition 2 and 3, that for all $t \in \mathbb{N}$,

$$E[|S_t|] \leq E[b + \varepsilon T_a] < \infty. \tag{18}$$

Also, by the definition of the process, for all $t \in \mathbb{N}$ it holds in the case $t \geq T_a$ that,

$$E\left[S_{t+1}\,; t \geq T_a \mid \mathscr{F}_t\right] = E\left[S_t\,; t \geq T_a \mid \mathscr{F}_t\right].$$

Furthermore, for all $t \in \mathbb{N}$, it holds in the case $t < T_a$,

$$\begin{aligned} &E\left[S_{t+1}\,; t < T_a \mid \mathscr{F}_t\right] \\ &\quad = E\left[(Z_{t+1} - Z_t + \varepsilon) + Z_t + \varepsilon t\,; t < T_a \mid \mathscr{F}_t\right] \\ &\quad \leq E\left[Z_t + \varepsilon t\,; t < T_a \mid \mathscr{F}_t\right] = E\left[S_t\,; t < T_a \mid \mathscr{F}_t\right], \end{aligned}$$

where the inequality is due to condition 1.1. Combining both cases, we have for all $t \in \mathbb{N}$,

$$E\left[S_{t+1} \mid \mathscr{F}_t\right] \leq E\left[S_t \mid \mathscr{F}_t\right] = S_t. \tag{19}$$

By (18) and (19), S_t is a super-martingale, thus for all $t \in \mathbb{N}$,

$$E\left[S_t \mid \mathscr{F}_0\right] \leq E\left[S_0 \mid \mathscr{F}_0\right] = Z_0. \tag{20}$$

By (17) and (18), the dominated convergence theorem applies (see 9.7 (g) in [57]), and by (20), $Z_0 \geq \lim_{t\to\infty} E\left[S_t \mid \mathscr{F}_0\right] = E\left[\lim_{t\to\infty} S_t \mid \mathscr{F}_0\right] = E\left[Z_{T_a} + \varepsilon T_a \mid \mathscr{F}_0\right]$. By noting that $Z_{T_a} \geq 0$, the upper bound proof is complete.

The lower bound proof follows exactly the same steps, i. e. starting by defining the same stopped process S_t. However, because the directions of the inequalities are inverted, S_t is a sub-martingale, and in the end we overestimate Z_{T_a} by a. □

We also state the lemmas used in the proof of Theorem 1. Their proofs are provided in the supplementary materials.

Lemma 18. *The functions g_1 and g_2 defined below are level functions for any $c > 0$, $\kappa \in (0,1)$, $x \in [\lambda]$, $y \in [m]$ and $\gamma_j, q_j \in (0,1]$ for each $j \in [m-1]$.*

$$g_1(x,y) := \ln\left(\frac{1+c\lambda}{1+c\max\{x,\gamma_y\lambda\}}\right) + \sum_{i=y+1}^{m-1} \ln\left(\frac{1+c\lambda}{1+c\gamma_i\lambda}\right),$$

$$g_2(x,y) := \frac{(1-\kappa)^x}{q_y} + \sum_{i=y+1}^{m-1} \frac{1}{q_i}$$

for all $y \in [m-1]$, and $g_1(x,y) := g_2(x,y) := 0$ for $y = m$.

Lemma 19 (Improved version of Lemma 5 in [14])**.** *If $X \sim \mathrm{Bin}(\lambda, p)$ with $p \geq (i/\lambda)(1+\delta)$ and $i \geq 1$ for some $\delta \in (0,1]$, then*

$$E\left[\ln\left(\frac{1+\delta X/2}{1+\delta i/2}\right)\right] \geq \frac{\delta^2}{7}.$$

Lemma 20. *Let $\{X_i\}_{i\in[\lambda]}$ be i.i.d. random variables, define $Y(j) := \sum_{i=1}^{\lambda} \mathbb{1}_{\{X_i \geq j\}}$ for any $j \in \mathbb{R}$. It holds for any $a, b, c, j \in \mathbb{R}$ with $c \geq 0$ and $b \leq \lambda$ that*

$$(i)\ \ \Pr\left(Y(j+c) \geq a \mid Y(j) \geq b\right) \geq \Pr\left(Y(j+c) \geq a\right)$$

and for any non-decreasing function f

$$(ii)\ \ E\left[f(Y(j+c)) \mid Y(j) \geq b\right] \geq E\left[f(Y(j+c))\right]$$

provided that both expectations are well-defined.

Appendix B Proofs Omitted in the Paper

Proof of Lemma 6. Let $f_1 \geq f_2 \geq \cdots \geq f_\lambda$ be the f-values of the individuals in the population P in decreasing order, and $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_\lambda$ be the levels of the individuals in the population P in decreasing order. Then, by the definition of the functions β and ζ, it suffices to prove that for all $j \in [\lambda]$, and all $x \in P$,

$$[x \in A_{\geq \ell_j}] \geq [f(x) \geq f_j].$$

Assume by contradiction that there exists an individual $x \in P \cap A_k$ such that $f_j \leq f(x)$ and $k < \ell_j$. Since by assumption $x \in A_k$ where $k < \ell_j$, the definition of f-based partitions imply that

$$f(x) < \min_{y \in A_{\ell_j}} f(y) = \min_{y \in A_{\geq \ell_j}} f(y). \tag{21}$$

Since f_j is the fitness of the j-th best individual ranked according to fitness

$$|\{z \in P \mid f(z) > f_j\}| \leq j - 1. \tag{22}$$

Thus, the assumption $f_j \leq f(x)$, inequality (21), and inequality (22) imply

$$|\{z \in P \mid f(z) \geq \min_{y \in A_{\geq \ell_j}} f(y)\}| \leq j - 1. \tag{23}$$

Since ℓ_j is the level of the j-th best individual ranked according to levels

$$|\{z \in P \mid z \in A_{\geq \ell_j}\}| \geq j. \tag{24}$$

By the definition of f-based partitions, it holds

$$\forall z \in A_{\geq \ell_j}, \quad f(z) \geq \min_{y \in A_{\geq \ell_j}} f(y). \tag{25}$$

Now inequalities (24) and (25) imply that

$$|\{z \in P \mid f(z) \geq \min_{y \in A_{\geq \ell_j}} f(y)\}| \geq j, \tag{26}$$

which contradicts with inequality (23). Thus we conclude that for all individuals $x \in P$ with fitness $f(x) \geq f_j$, it holds $x \in A_{\geq \ell_j}$, which is equivalent to the statement $[x \in A_{\geq \ell_j}] \geq [f(x) \geq f_j]$. □

Proof of Lemma 9. Since (M3) only concerns the restricted subspace $\mathcal{X} \setminus A_m$ we only need to focus on this subspace. Furthermore, since the partition is f-based on this subspace, it suffices by Lemma 6 to prove the results for the β-function instead of the ζ-function.

The results for k-tournament selection, (μ, λ)-selection, and linear ranking selection follow by applying Lemma 13 in [14] (with its p_0 being set as our $p_0(1 - p_c)$). For exponential ranking, we first note the following lower bound,

$$\begin{aligned}\beta(\gamma, P) &\geq \int_0^\gamma \frac{\eta e^{\eta(1-x)} dx}{e^\eta - 1} = \left(\frac{e^\eta}{e^\eta - 1}\right)\left(1 - \frac{1}{e^{\eta\gamma}}\right) \\ &\geq 1 - \frac{1}{1 + \eta\gamma}.\end{aligned}$$

The rest of the proof is similar to that for k-tournament selection with η in place of k. □

Proof of Lemma 10. Similarly to the proof of Lemma 9, we only focus on the subspace $\mathcal{X} \setminus A_m$ where the partition is f-based. Using Lemma 6, we consider the β-function instead of the ζ-function. First define $\varepsilon' := \varepsilon p_0$.

1. Consider k-tournament selection and let $\gamma \in (0, \gamma_0]$. By the definition of f-based partitions, to select an individual from the $\gamma\lambda$ best individuals it is sufficient that the randomly sampled tournament contains at least one of the $\gamma\lambda$ best individuals. Hence,

$$\beta(\gamma, P) \geq 1 - (1-\gamma)^k > 1 - \frac{1}{1+\gamma k},$$

because $(1-\gamma)^k < e^{-\gamma k} < \frac{1}{1+\gamma k}$. So for $k \geq 4(1+\delta')/\varepsilon'$,

$$\beta(\gamma, P) \geq 1 - \frac{1}{1+\frac{4\gamma(1+\delta')}{\varepsilon'}} = \frac{\frac{4\gamma(1+\delta')}{\varepsilon'}}{1+\frac{4\gamma(1+\delta')}{\varepsilon'}}.$$

If $\gamma_0 := \varepsilon'/(4(1+\delta'))$, then for all $\gamma \in (0, \gamma_0]$ it holds that $4(1+\delta')/\varepsilon' \leq 1/\gamma$ and

$$\begin{aligned}\beta(\gamma, P) &\geq \frac{\gamma 4(1+\delta')/\varepsilon'}{\gamma(1/\gamma)+1} = \frac{2(1+\delta')\gamma}{\varepsilon'} \\ &= \sqrt{\frac{(1+\delta')}{\varepsilon'(\varepsilon'/4(1+\delta'))}}\gamma = \sqrt{\frac{(1+\delta')}{\varepsilon'\gamma_0}}\gamma.\end{aligned}$$

2. In (μ, λ)-selection, again by the f-based property of the partition, we have $\beta(\gamma, P) = \lambda\gamma/\mu$ if $\gamma\lambda \leq \mu$, and $\beta(\gamma, P) = 1$ otherwise. It suffices to pick $\gamma_0 := \mu/\lambda$ so that with $\lambda/\mu \geq (1+\delta')/\varepsilon'$, for all $\gamma \in (0, \gamma_0]$. Then

$$\beta(\gamma, P) \geq \frac{\lambda\gamma}{\mu} = \sqrt{\frac{\lambda^2}{\mu^2}}\gamma = \sqrt{\frac{\lambda}{\mu\gamma_0}}\gamma \geq \sqrt{\frac{1+\delta'}{\varepsilon'\gamma_0}}\gamma.$$

3. Similarly to the proof of Lemma 9, we remark that $\beta(\gamma, P) \geq 1 - \frac{1}{1+\eta\gamma}$, thus the rest of the proof is similar to k-tournament selection. □

Proof of Theorem 11. We apply Corollary 7 with the canonical partition $A_j := \{j \mid f(x) = j\}$ for all functions[1], except for LINEAR, where the partition from [14] is used: $A_n := \{1^n\}$, $A_j := \left\{x \mid \sum_{i=1}^{j} c_i \leq f(x) < \sum_{i=1}^{j+1} c_i\right\}$ for $j \in \{0\} \cup [n-1]$.

The choices of s_j and s_* to satisfy (M1) are the following.

- For ONEMAX, $s_j := \binom{n-j}{1}\left(\frac{\chi}{n}\right)\left(1-\frac{\chi}{n}\right)^{n-1} = \Omega\left(\frac{n-j}{n}\right)$, i. e. the probability of flipping a 0-bit while keeping all the other bits unchanged, and $s_* := \Omega\left(\frac{1}{n}\right)$.
- For LEADINGONES, ℓ-UNIMODAL and LINEAR, $s_j := \frac{\chi}{n}\left(1-\frac{\chi}{n}\right)^{n-1} = \Omega\left(\frac{1}{n}\right) =: s_*$, i. e. the probability of flipping a specific bit to create a Hamming neighbour solution with better fitness while keeping all the other bits

[1] The first level can be A_0 instead of A_1 for some functions but that does not matter as far as we compute the sums correctly later on.

unchanged. In ℓ-UNIMODAL, the bit to flip must exist by the definition of the function. In LEADINGONES, the 0-bit at position $j+1$ should be flipped. For LINEAR, the partition satisfies that among the first $j+1$ bits, there must be at least a 0-bit, thus it suffices to flip the left most 0-bit will produce a search point at a higher level.

- For JUMP_r, similarly to ONEMAX for $j \in [n-1]$

$$s_j := \binom{n-j+1}{1}\left(\frac{\chi}{n}\right)\left(1-\frac{\chi}{n}\right)^{n-1} = \Omega\left(\frac{n-j+1}{n}\right),$$

but $s_n := \left(\frac{\chi}{n}\right)^r\left(1-\frac{\chi}{n}\right)^{n-r} = \Omega\left(\left(\frac{\chi}{n}\right)^r\right)$, i.e. the probability of flipping the r remained 0-bits, so $s_* := \Omega\left(\left(\frac{1}{n}\right)^r\right)$.

- For ROYALROAD_r, $s_j := \binom{n/r-j}{1}\left(\frac{\chi}{n}\right)^r\left(1-\frac{\chi}{n}\right)^{n-r} = \Omega\left(\left(\frac{\chi}{n}\right)^r\left(\frac{n}{r}-j\right)\right)$, i.e. the probability of flipping an entire unsolved block of length r (in the worst case) while keeping the other bits unchanged, and $s_* := \Omega\left(\left(\frac{1}{n}\right)^r\right)$.

Lemma 29 with $\varepsilon = \frac{\delta/2}{1+\delta/2}$ implies that the probability of not flipping any bit position by mutation is $(1-\chi/n)^n \geq \left(1-\frac{\delta/2}{1+\delta/2}\right)e^{-\chi} = \frac{e^{-\chi}}{1+\delta/2}$ for n sufficiently large. Thus choosing $p_0 := \frac{e^{-\chi}}{1+\delta/2}$ satisfies (M2).

We now look at (M3). For k-tournament selection, we have

$$k \geq \frac{(1+\delta)e^{\chi}}{1-p_c} = \left(1+\frac{\delta/2}{1+\delta/2}\right)\frac{1}{(1-p_c)p_0}$$

due to our choice of p_0. Hence, it follows from Lemma 9 that (M3) is satisfied with constant $\delta' := \frac{\delta/2}{1+\delta/2}$. The same conclusion can be drawn for the other three selection mechanisms.

In (M4), since γ_0 and δ' are constants, there should exist a constant $c > 0$ for each function such that the condition is satisfied given the minimum requirement on population size related to c.

Since all conditions are satisfied, Corollary 7 gives the desired result for each function. For ONEMAX and JUMP_r, optimisation time can be saved at early levels, thus the evaluation of the sum $\sum_{j=1}^{m-1}\ln\left(\frac{6\delta\lambda}{4+\gamma_0 s_j \delta\lambda}\right)$ has to be precise:

- For ONEMAX, bounding each term by $\ln\left(\frac{6}{\gamma_0 s_j}\right)$ gives

$$\mathcal{O}\left(\ln\left(\frac{6^n n^n}{\gamma_0^n \prod_{j=0}^{n-1}(n-j)}\right)\right)$$

and by Stirling's formula $\prod_{j=0}^{n-1}(n-j) = n! = \Theta(n^{n+\frac{1}{2}}/e^n)$, so this sum is no more than $\mathcal{O}(n)$.

- For JUMP_r, using the upper bound $\ln\left(\frac{6}{\gamma_0 s_j}\right)$ for the first $m-2$ terms of the sum, and the upper bound $\ln(3\lambda/2)$ for the last term gives the upper bound $\mathcal{O}(n+\ln\lambda)$ for this sum.

For the other functions, we bound the sum by $\mathcal{O}(m \ln \lambda)$. The evaluation of the other sum $\sum_{j=1}^{m-1} 1/s_j$ is standard to the fitness level technique [56]. Take OneMax as an example, $\sum_{j=1}^{m-1} 1/s_j = \mathcal{O}\left(\sum_{j=0}^{n-1} \frac{n}{n-j}\right) = \mathcal{O}(nH_n) = \mathcal{O}(n \ln n)$ where H_n is the n-th harmonic number. Thus combining the two sums gives that $E[T] = \mathcal{O}(n\lambda + n \ln n) = \mathcal{O}(n\lambda)$ (since $\lambda = \Omega(\ln n)$) for this function. This can be done analogously for the other functions, and the results follow. □

Proof of Theorem 12. This time we apply Corollary 8 , again using the canonical partition of the search space for both functions. We also assume that n is large enough so that by Lemma 29 the probability of not flipping any bit by mutation is $(1-\chi/n)^n \geq \left(1 - \frac{\delta/2}{1+\delta/2}\right) e^{-\chi} = \frac{e^{-\chi}}{1+\delta/2} =: p_0$, and so (C2) is satisfied with this choice of p_0. In addition, we use the same upgrade probabilities s_j and their smallest value s_* for each of the two functions as in the proof of Theorem 11 to satisfy (C1).

It follows from Lemma 23 (see the next section) that (C3) is satisfied for constant $\varepsilon_1 := 1/2$. We now look at condition (C4). For k-tournament, we get $k \geq 8(1+\delta)e^{\chi} = 4\left(1 + \frac{\delta/2}{1+\delta/2}\right)/(p_0\varepsilon_1)$. So condition (C4) is satisfied with constant $\delta' := \frac{\delta/2}{1+\delta/2}$ for k-tournament by Lemma 10. The same reasoning can be applied so that (C4) is also satisfied for the other selection mechanisms.

Since δ' and γ_0 are constants, thus condition (C5) is satisfied given $\lambda \geq c \ln n$ and for some constant c. Since all conditions are satisfied, the result follows from Corollary 8. □

Proof of Theorem 13. Define $m := \binom{n}{2}$. We apply Corollary 7 with the canonical partition, $A_j := \{\pi \mid \textsc{Inv}(\pi) = j\}$ for $j = \{0\} \cup [m]$. The probability that mutation exchanges 0 pairs is $1/e$. Hence, condition (M2) is satisfied for $p_0 := 1/e$.

To show that (M1) is satisfied, we first define $s_j := \frac{m-j}{em}$ for each $j \in \{0\} \cup [m-1]$. Since $x \in A_j$, then the probability that the exchange operator exchanges exactly one pair is $1/e$, and the probability that this pair is incorrectly ordered in x, is $(m-j)/m$. Thus, (M1) is satisfied with the defined s_j.

In (M3), for k-tournament we have that $k \geq \frac{(1+\delta)e}{1-p_c} = \left(1 + \frac{\delta/2}{1+\delta/2}\right) \frac{1}{(1-p_c)p_0}$, thus the condition is satisfied for constant $\delta' := \frac{\delta/2}{1+\delta/2}$ and some constant $\gamma_0 \in (0,1)$ by Lemma 9. The same conclusion can be drawn for the other selection mechanisms. Finally, since γ_0, δ' are constants, there exists a constant $c > 0$ such that (M4) is satisfied for any $\lambda \geq c \ln(n)$.

It therefore follows by Corollary 7 that the expected runtime of the GA on $\textsc{Sorting}_{\textsc{Inv}}$ is $\mathcal{O}\left(n^2\lambda\right)$, i. e. this is similar to OneMax except that we have $m = O(n^2)$ levels. □

Proof of Corollary 14. We use the following partition

$$A_j := \{x \in \mathcal{X} | f(x) = f_j\} \backslash \mathcal{LO},\ j \in [m-1], \text{ and}$$
$$A_m := \mathcal{LO}.$$

We note that $(A_1, \ldots, A_{m-1})$ is a fitness-based partition of $\mathcal{X} \setminus \mathcal{LO}$. Thus, applications of Corollary 7 and Lemma 9 for the set of conditions (X1–2) and (X4.2), or alternatively, Corollary 8 and Lemma 10 for the set of conditions (X1–3) and (X4.1), yield the required result. □

Proof of Theorem 16. Step 1: We use the canonical partition into $m = n + 1$ levels, where level $j \in [m]$ is defined by

$$A_j := \{x \in \{0,1\}^n \mid \textsc{OneMax}(x) = j - 1\}.$$

We use the parameter $\gamma_0 := \mu/\lambda$ and let Y be the number of one-bits in a sampled solution from p_t.

The choice $m' = \mu/n$ and $\mu \leq n$ implies that the margins for $p_t(i)$ are simplified to $1/n$ and $1 - 1/n$, and that these margins are only used when the bit values at position i of the μ selected individuals are identical. We categorise the probabilities $p_t(i)$ into three groups: those at the upper margin $1 - 1/n$, those at the lower margin $1/n$, and intermediary values in the closed interval $[1/\mu, 1-1/\mu]$. Due to linearity of the fitness function, the components of p_t can be rearranged without changing the distribution of Y. We assume w.l.o.g. a rearrangement so that there exist integers $k, \ell \geq 0$ satisfying

$$\begin{aligned} 1 \leq X_i < \mu \quad &\text{and} \quad p_t(i) = X_i/\mu && \text{if } 1 \leq i \leq k, \\ X_i = \mu \quad &\text{and} \quad p_t(i) = 1 - 1/n && \text{if } k < i \leq k + \ell, \text{and} \\ X_i = 0 \quad &\text{and} \quad p_t(i) = 1/n && \text{if } k + \ell < i \leq n. \end{aligned}$$

By these assumptions, it follows that

$$\sum_{i=k+1}^{k+\ell} X_i = \mu\ell \quad \text{and} \quad \sum_{i=k+\ell+1}^{n} X_i = 0. \tag{27}$$

In the following, we define $Y_{i,k}$ to be the number of sampled one-bits due to $(p_t(i), \ldots, p_t(k))$ in the rearranged p_t.

For any population P_t and any $\gamma \in [0, \gamma_0]$, let $j \in [n]$ be any integer such that $|P_t \cap A_{\geq j}| \geq \gamma_0\lambda = \mu$ and $|P_t \cap A_{\geq j+1}| \geq \gamma\lambda$. This implies that among the μ fittest individuals in the current population, there are at least $\gamma\lambda$ individuals with at least j one-bits, and the remaining among the μ fittest individuals have at least $j - 1$ one-bits. Hence, the total number of one-bits among the fittest μ individuals must satisfy

$$\sum_{i=1}^{n} X_i \geq \gamma\lambda j + (\mu - \gamma\lambda)(j-1) = \gamma\lambda + \mu(j-1). \tag{28}$$

Combining Eqs. (27) and (28), when $k \geq 1$, we get

$$E[Y_{1,k}] = \sum_{i=1}^{k} p_t(i) = \frac{1}{\mu} \sum_{i=1}^{k} X_i \geq \frac{\gamma\lambda}{\mu} + j - 1 - \ell. \tag{29}$$

Step 2: We first verify condition (G2), i.e. checking if $\Pr(Y \geq j) \geq (1+\delta)\gamma$ for any level j defined like above with $\gamma > 0$. We distinguish between two cases, either $k = 0$ or $k \geq 1$.

Case 1: If $k \geq 1$, then Eq. (29) and Lemma 30 give

$$\begin{aligned} \Pr(Y_{1,k} \geq j - \ell) &= \Pr\left(Y_{1,k} > j - 1 - \ell + \frac{\gamma\lambda}{\mu} - \frac{\gamma\lambda}{12\mu}\right) \\ &\geq \Pr\left(Y_{1,k} > E[Y_{1,k}] - \frac{\gamma\lambda}{12\mu}\right) \end{aligned}$$

$$\geq \min\left\{\frac{1}{13}, \frac{\frac{\gamma\lambda}{12\mu}}{\frac{\gamma\lambda}{12\mu}+1}\right\}$$
$$= \min\left\{\frac{1}{13}, \frac{\gamma\lambda}{\gamma\lambda+12\mu}\right\} \geq \frac{\gamma\lambda}{13\mu}.$$

The probability of sampling an individual with at least j one-bits in the next generation is therefore lower-bounded by

$$\begin{aligned}\Pr(Y \geq j) &\geq \Pr(Y_{1,k} \geq j-\ell)\Pr(Y_{k+1,k+\ell} = \ell)\\ &\geq \frac{\gamma\lambda}{13\mu}\left(1-\frac{1}{n}\right)^{\ell} \geq \frac{\gamma\lambda}{13\mu}\left(1-\frac{1}{n}\right)^{n-1}\\ &\geq \frac{\gamma\lambda}{13e\mu} \geq (1+\delta)\gamma.\end{aligned}$$

Case 2: If $k = 0$, then all the μ best individuals in the population must be identical. By assumption, there are $\gamma\lambda \geq 1$ individuals with at least j 1-bits, hence all the μ best individuals must have at least j 1-bits. In this case, there are $\ell \geq j$ probabilities at the upper margin, and we get

$$\begin{aligned}\Pr(Y \geq j) &\geq \Pr\left(Y_{k+1}^{k+j} \geq j\right)\\ &= \left(1-\frac{1}{n}\right)^{j} \geq \frac{1}{e} \geq 13\gamma_0(1+\delta) > (1+\delta)\gamma,\end{aligned}$$

and condition (G2) is therefore satisfied also in this case.

Step 3: We now consider condition (G1) for any j defined. Again we check the two cases $k = 0$, and $k \geq 1$.

Case 1: If $k = 0$, then with our assumption, the $\ell \geq j-1$ first probabilities are at the upper margin $1-1/n$, and the last $n-\ell \leq n-j+1$ probabilities are at the lower margin $1/n$. In order to obtain a search point with at least j one-bits, it is sufficient to sample exactly $\ell \geq j-1$ one-bits in the first ℓ positions and exactly one 1-bit in the last $n-\ell \leq n-j+1$ positions. Hence,

$$\begin{aligned}\Pr(Y \geq j) &\geq \Pr(Y_{1,\ell} \geq \ell)\Pr(Y_{\ell+1,n} \geq 1)\\ &\geq \left(1-\frac{1}{n}\right)^{\ell}\left(\frac{n-\ell}{n}\right)\\ &\geq \left(1-\frac{1}{n}\right)^{n-1}\left(\frac{n-j+1}{n}\right) \geq \left(\frac{n-j+1}{en}\right).\end{aligned}$$

Case 2: When $k \geq 1$, we note from Eq. (29) that

$$E[Y_{2,k}] = E[Y_{1,k}] - p_t(1) \geq j-1-\ell-p_t(1)$$

Again, by Lemma 30, we get

$$\begin{aligned}&\Pr(Y_{2,k} \geq j-1-\ell)\\ &\quad= \Pr(Y_{2,k} > j-1-\ell-p_t(1)-(1-p_t(1)))\\ &\quad\geq \Pr(Y_{2,k} > E[Y_{2,k}]-(1-p_t(1)))\end{aligned}$$

$$\geq \min\left\{\frac{1}{13}, \frac{1-p_t(1)}{2-p_t(1)}\right\} > \frac{1-p_t(1)}{13}.$$

The probability of sampling an individual with at least j one-bits in this configuration is bounded from below as

$$\begin{aligned}
&\Pr(Y \geq j) \\
&\qquad > \Pr(Y_1 = 1) \Pr(Y_{2,k} \geq j-1-\ell) \Pr(Y_{k+1,k+\ell} = \ell) \\
&\qquad\qquad \geq \left(\frac{p_t(1)(1-p_t(1))}{13}\right)\left(1-\frac{1}{n}\right)^{\ell} \\
&\qquad\qquad\qquad \geq \left(\frac{(1/\mu)(1-1/\mu)}{13}\right)\left(1-\frac{1}{n}\right)^{\ell} \geq \frac{1}{14e\mu}.
\end{aligned}$$

The last inequality holds for $\mu \geq 14$, which in turn only requires n to be larger than some constant. Hence, combining the cases $k = 0$ and $k > 0$, we get for all $j \in [n]$,

$$\begin{aligned}
\Pr(Y \geq j) &\geq \min\left\{\frac{1}{14e\mu}, \frac{n-j+1}{en}\right\} \\
&\geq \frac{n-j+1}{14e\mu(n-j+1)+en} =: z_j.
\end{aligned}$$

Clearly, there exists a z_* with $1/z_* \in \text{poly}(n)$ such that $\Pr(Y \geq j) \geq z_*$ for all $j \in [n]$ and condition (G1) is satisfied.

Step 4: We consider condition (G3) regarding the population size. The parameters δ and $\gamma_0 = \mu/\lambda$ are constants with respect to n, therefore the variables a, ε and c in condition (G3) are also constants, and $1/z_* \in \text{poly}(n)$. Hence, there must exist a constant $b > 0$ such that condition (G3) is satisfied when $\lambda \geq b\log(n)$.

Step 5: To conclude, the expected optimisation time is

$$\begin{aligned}
&\mathcal{O}\left(n\lambda\ln(\lambda) + \sum_{j=1}^{n}\frac{1}{z_j}\right) \\
&= \mathcal{O}\left(n\lambda\ln(\lambda) + 14e\mu n + \sum_{i=0}^{n-1}\frac{en}{n-i}\right) = \mathcal{O}\left(n\lambda\ln\lambda\right). \qquad \square
\end{aligned}$$

Proof of Theorem 17. We construct an operator D which leads to the claimed lower bound. Choose any sequence of search points $(x_1, \ldots, x_m) \in A_1 \times \cdots \times A_m$, and let the initial population of Algorithm 1 be $P_0 := (x_1, \ldots, x_1)$, i. e., λ copies of the search point x_1 belonging to the first level.

For any population $P \in \mathcal{X}^\lambda$, let the *current level* be the largest $i \in [m]$ such that $|P \cap A_{\geq i}| \geq \gamma_0\lambda$. For any population P with current level $i < m$, define the operator D for all $u \in \mathcal{X}$ by

$$\Pr_{y\sim D(P)}(y = u) := \begin{cases} 1 - \max\{(1+\delta)\gamma, z_i\} & \text{if } u = x_i \\ \max\{(1+\delta)\gamma, z_i\} & \text{if } u = x_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \tag{30}$$

where $\gamma := (1/\lambda)|P \cap A_{\geq i+1}| < \gamma_0$.

For all $t \in \mathbb{N}$, define

$$T_j := \min\{t \mid |P_t \cap A_{\geq j}| > 0\}, \text{ for all } j \in [m], \text{ and}$$
$$S_j := T_{j+1} - T_j \text{ for all } j \in [m-1].$$

Then we have $\sum_{j=1}^{m-1} S_j = T_m - T_1 = T$ because $T_1 = 0$. The random variable S_j, for $j \in [m-1]$, describes the number of generations from the time the process has discovered the search point x_j until it has discovered the search point x_{j+1}, and we call this *phase* j. We divide each phase j into two sub-phases. Let S_j^1 be the number of generations, where

$$1 \leq |P_t \cap A_{\geq j}| < \gamma_0\lambda,$$

and call this the first sub-phase, and let S_j^2 be the number of generations, where

$$\gamma_0\lambda \leq |P_t \cap A_{\geq j}| \text{ and } 0 = |P_t \cap A_{\geq j+1}|,$$

and call this the second sub-phase. The duration of the j-th phase is the sum $S_j = S_j^1 + S_j^2$. Remark that $S_1^1 = 0$ due to the choice of the initial population P_0.

Note also that by the definition of operator D, as long as the process is in sub-phase 1 of phase j, the probability of generating the search point x_{j+1} is 0. Furthermore, the process never returns to sub-phase 1 once the process has entered sub-phase 2. To estimate the duration of sub-phase 1, we consider the stochastic process $(X_t)_{t\in\mathbb{N}}$, where $X_t := |P_{T_j+t} \cap A_{\geq j}|$, and a corresponding filtration $(\mathscr{F}_t)_{t\in\mathbb{N}}$, where $\mathscr{F}_t := \sigma\left(P_1, \ldots, P_{T_j+t}\right)$.

During sub-phase 1 of phase $j > 1$, it holds that $X_{t+1} \sim \text{Bin}(\lambda, p_{t+1})$, where $p_{t+1} = \max\left\{(1+\delta)\frac{X_t}{\lambda}, z_{j-1}\right\}$.

To lower bound the expected duration of sub-phase 1, we apply drift analysis (Theorem 3) with respect to the process $(Z_t)_{t\in\mathbb{N}}$ defined by $Z_t := \ln(\lambda/R_t)$, where $R_t := \max\{X_t, y_j\}$ and $y_j := \max\{\lambda z_{j-1}, 1/\delta^2\} > 1$. Note that since $z_j < \gamma_0$ by assumption, and $1/\delta^2 < \gamma_0\lambda$ by condition (G3), it holds that $y_j < \gamma_0\lambda$. It is therefore clear that sub-phase 1 is only complete if

$$Z_t \leq \ln\left(\frac{\lambda}{\gamma_0\lambda}\right) = -\ln(\gamma_0) =: a.$$

By Jensen's inequality, the drift of this process can be bounded by

$$\begin{aligned} E\left[Z_t - Z_{t+1} \mid \mathscr{F}_t\right] &= E\left[\ln\left(\frac{R_{t+1}}{R_t}\right) \mid \mathscr{F}_t\right] \\ &\leq \ln\left(\frac{E\left[R_{t+1} \mid \mathscr{F}_t\right]}{R_t}\right) \end{aligned}$$

and by Lemma 24

$$\leq \ln\left(\frac{\max(\lambda p_{t+1}, y_j) + (\frac{1}{2})\sqrt{\lambda p_{t+1}}}{R_t}\right).$$

We consider two sub-cases. Either $\lambda p_{t+1} \geq y_j$, in which case we use that

$$\begin{aligned} R_t &= \max\{X_t, y_j\} \\ &= \max\{X_t, \lambda z_{j-1}, 1/\delta^2\} \\ &\geq \lambda p_{t+1}/(1+\delta) \end{aligned}$$

because $p_{t+1} = \max\{X_t(1+\delta)/\lambda, z_{j-1}\}$, so

$$\begin{aligned} E\left[Z_t - Z_{t+1} \mid \mathscr{F}_t\right] &\leq \ln\left(\frac{\lambda p_{t+1} + (1/2)\sqrt{\lambda p_{t+1}}}{\lambda p_{t+1}/(1+\delta)}\right) \\ &= \ln\left((1+\delta)\left(1 + \frac{1}{2\sqrt{\lambda p_{t+1}}}\right)\right). \end{aligned}$$

From the assumption $\lambda p_{t+1} \geq y_j$ and $y_j > 1$, it follows that $\sqrt{1/(\lambda p_{t+1})} \leq \sqrt{1/y_j} \leq \delta$ and

$$E\left[Z_t - Z_{t+1} \mid \mathscr{F}_t\right] \leq \ln(1+\delta) + \ln(1+\delta/2) < (3/2)\delta.$$

In the other sub-case, when $y_j > \lambda p_{t+1}$, we use that $R_t \geq y_j$ and get

$$\begin{aligned} E\left[Z_t - Z_{t+1} \mid \mathscr{F}_t\right] &\leq \ln\left(\frac{y_j + (1/2)\sqrt{\lambda p_{t+1}}}{y_j}\right) \\ &< \ln\left(1 + \frac{\sqrt{y_j}}{2y_j}\right) = \ln\left(1 + \frac{1}{2\sqrt{y_j}}\right) \\ &\leq \ln\left(1 + \frac{\delta}{2}\right) < \delta/2. \end{aligned}$$

Hence, condition 1 in Theorem 3 can be satisfied with the parameter $\varepsilon := (3/2)\delta$. We therefore get the bound

$$E\left[S_j^1 \mid \mathscr{F}_0\right] \geq \frac{Z_0 - a}{\varepsilon} = \left(\frac{2}{3\delta}\right)\ln\left(\frac{\gamma_0\lambda}{\max\{X_0, \lambda z_{j-1}, \frac{1}{\delta^2}\}}\right).$$

By the definition of the process, for $1 < j \leq m$, we have $X_0 \sim (Y \mid Y \geq 1)$, where $Y \sim \mathrm{Bin}(\lambda, z_{j-1})$, i. e., X_0 is random variable with distribution equal to a binomial distribution conditioned on taking value at least 1. By the tower property of expectation,

$$\begin{aligned} E\left[S_j^1\right] &= E\left[E\left[S_j^1 \mid \mathscr{F}_0\right]\right] \\ &\geq E\left[\left(\frac{2}{3\delta}\right)\ln\left(\frac{\gamma_0\lambda}{\max\{X_0, \lambda z_{j-1}, 1/\delta^2\}}\right)\right] \\ &> E\left[\left(\frac{2}{3\delta}\right)\ln\left(\frac{\gamma_0\lambda}{X_0 + \lambda z_{j-1} + 1/\delta^2}\right)\right], \\ &= E\left[\left(\frac{2}{3\delta}\right)\ln\left(\frac{\gamma_0\lambda}{Y + \lambda z_{j-1} + 1/\delta^2}\right) \mid Y \geq 1\right], \end{aligned}$$

since the function $f(x) = \ln(1/x)$ is convex, Jensen's inequality and Lemma 25 give

$$> \left(\frac{2}{3\delta}\right)\ln\left(\frac{\gamma_0\lambda}{E\left[Y \mid Y \geq 1\right] + \lambda z_{j-1} + 1/\delta^2}\right)$$

$$\geq \left(\frac{2}{3\delta}\right) \ln\left(\frac{\gamma_0 \lambda}{1 + 2\lambda z_{j-1} + 1/\delta^2}\right).$$

During sub-phase 2, it holds that

$$\Pr_{y \sim D(P_t)}(y = x_j) = 1 - z_j, \text{ and } \Pr_{y \sim D(P_t)}(y = x_{j+1}) = z_j.$$

In each generation of sub-phase 2, the phase ends with probability $q_j := 1 - (1 - z_j)^\lambda < \lambda z_j$, i. e., the probability that at least one individual is produced in $A_{\geq j+1}$. The duration of sub-phase 2 is therefore geometrically distributed with parameter q_j and has expectation $E\left[S_j^2\right] = 1/q_j \geq 1/(\lambda z_j)$.

Hence, we get

$$\begin{aligned} E[T] &= \sum_{j=1}^{m-1} E\left[S_j^1\right] + E\left[S_j^2\right] \\ &\geq \left(\frac{2}{3\delta} \sum_{j=1}^{m-2} \ln\left(\frac{\gamma_0 \lambda}{1 + 2\lambda z_j + 1/\delta^2}\right)\right) + \sum_{j=1}^{m-1} \frac{1}{\lambda z_j}. \end{aligned}$$ □

Appendix C Proofs Omitted from the Appendix

Proof of Lemma 18. Both g_1 and g_2 are non-increasing functions in x and y, hence properties 1 and 2 of Definition 4 are satisfied. Property 3 is satisfied because for all $y \in [m-1]$

$$\begin{aligned} g_1(\lambda, y) &= \sum_{i=y+1}^{m-1} \ln\left(\frac{1 + c\lambda}{1 + c\gamma_i \lambda}\right) \\ &= \ln\left(\frac{1 + c\lambda}{1 + c\gamma_{y+1}\lambda}\right) + \sum_{i=y+2}^{m-1} \ln\left(\frac{1 + c\lambda}{1 + c\gamma_i \lambda}\right) \\ &= g_1(0, y+1) \end{aligned}$$

and

$$\begin{aligned} g_2(\lambda, y) &= \frac{(1-\kappa)^\lambda}{q_y} + \sum_{i=y+1}^{m-1} \frac{1}{q_i} > \sum_{i=y+1}^{m-1} \frac{1}{q_i} \\ &= \frac{(1-\kappa)^0}{q_{y+1}} + \sum_{i=y+2}^{m-1} \frac{1}{q_i} = g_2(0, y+1). \end{aligned}$$ □

Proof of Lemma 19. Let $Y \sim \mathrm{Bin}(\lambda, (1+\delta)i/\lambda)$, then $Y \preceq X$. Therefore,

$$E\left[\ln\left(\frac{1 + \delta X/2}{1 + \delta i/2}\right)\right] \geq E\left[\ln\left(\frac{1 + \delta Y/2}{1 + \delta i/2}\right)\right]$$

and it is sufficient to show that $E\left[\ln\left(\frac{1 + \delta Y/2}{1 + \delta i/2}\right)\right] > \delta^2/7$ to complete the proof.

It follows from Corollary 22 (see the next section, and choose $c = \delta/2$) that

$$\begin{aligned}
&E\left[\ln\left(\frac{1+\delta Y/2}{1+\delta i/2}\right)\right] \\
&\geq \ln\left(\frac{1+(1+\delta)\delta i/2}{1+\delta i/2}\right) - \frac{\delta}{4}\cdot\frac{(1+\delta)\delta i/2}{1+(1+\delta)\delta i/2} \\
&= \ln\left(1+\frac{i\delta^2}{2+i\delta}\right) - \frac{\delta}{4}\cdot\frac{(1+\delta)\delta i}{2+(1+\delta)\delta i} =: h(i).
\end{aligned}$$

For all $\delta > 0$ and $i \geq 1$, it holds that

$$h'(i) = \frac{1}{2}\cdot\frac{(6+\delta(3i-2)+3i\delta^2)\delta^2}{(2+i\delta+i\delta^2)^2(2+i\delta)} > 0,$$

or $h(i)$ monotonically increases in i.

Define $r(\delta) := 12+8\delta+3\delta^2+\delta^3-2\delta^4$ and $s(\delta) := (2+\delta)^2(2+\delta+\delta^2) > 0$, we get

$$\begin{aligned}
h(i) \geq h(1) &= \ln\left(1+\frac{\delta^2}{2+\delta}\right) - \frac{\delta}{4}\cdot\frac{(1+\delta)\delta}{2+(1+\delta)\delta} \\
&\geq \frac{\delta^2}{2+\delta}\left(1-\frac{\delta^2}{2(2+\delta)}\right) - \frac{\delta}{4}\cdot\frac{(1+\delta)\delta}{2+(1+\delta)\delta} = \frac{\delta^2 r(\delta)}{4s(\delta)}.
\end{aligned}$$

The last inequality is due to Lemma 27. We notice that $18r(\delta) - 11s(\delta) = (1-\delta)(128+140\delta+84\delta^2+47\delta^3) \geq 0$ for all $\delta \in (0,1]$, thus $r(\delta)/s(\delta) \geq 11/18$ and $h(i) \geq (\delta^2/4)(11/18) > \delta^2/7$. □

Proof of Lemma 20. Define $p := \Pr(X_i \geq j)$ and $q := \Pr(X_i \geq j+c)$. For $b \leq 0$ or $p = 0$, the result trivially holds. For $b \in (0,\lambda]$ and $p \in (0,1]$, we have that $q' := \Pr(X_i \geq j+c \mid X_i \geq j) = q/p \geq q$. Event $Y(j) \geq b$ implies the existence of a set $A \subseteq [\lambda]$ such that $|A| \geq \lceil b \rceil$ and $X_i \geq j$ for all $i \in A$. Define $Y_1 := \sum_{i\in A} \mathbb{1}_{\{X_i \geq j+c\}}$ and $Y_2 := \sum_{i\in[\lambda]\setminus A} \mathbb{1}_{\{X_i\geq j+c\}}$, so $Y(j+c) = Y_1 + Y_2$. Clearly, conditioned on $Y(j) \geq b$, $Y_1 \sim \mathrm{Bin}(|A|, q') \succeq \mathrm{Bin}(|A|, q)$ and $Y_2 \sim \mathrm{Bin}(\lambda - |A|, q)$. Therefore, the distribution of $Y(j+c)$ conditioned on $Y(j) \geq b$ stochastically dominates $\mathrm{Bin}(|A|,q) + \mathrm{Bin}(\lambda-|A|,q) = \mathrm{Bin}(\lambda,q)$, which is the (unconditional or original) distribution of $Y(j+c)$, and part (i) follows.

For part (ii), let $F_1(x) := \Pr(f(Y(j+c)) < x \mid Y(j) \geq b)$ and $F_2(x) := \Pr(f(Y(j+c)) < x)$, i. e. F_1 and F_2 are the conditional and the unconditional distribution functions of $f(Y(j+c))$ respectively. Then from part (i), we conclude that $F_1(x) \leq F_2(x)$ for any $x \in \mathbb{R}$, and by the properties of expectation,

$$\begin{aligned}
E[f(Y(j+c)) \mid Y(j) \geq b] &= -\int_{-\infty}^{0} F_1(x)dx + \int_0^{\infty}(1-F_1(x))dx \\
&\geq -\int_{-\infty}^{0} F_2(x)dx + \int_0^{\infty}(1-F_2(x))dx \\
&= E[f(Y(j+c))].
\end{aligned}$$

□

Appendix D Lemmas Used in the Preceeding Proofs

The improvement achieved in Lemma 19 is due to the following generalisation of the lower bound of Lemma 33 in [14] (or Lemma 27 in this document).

Lemma 21. *For any $z > 0$, and all $x \geq 0$, we have that*

$$\ln(1+x) \geq x(b(z) + a(z)x),$$
$$\textit{where } a(z) := \frac{1}{z(z+1)} - \frac{\ln(1+z)}{z^2},$$
$$\textit{and } b(z) := \frac{2\ln(1+z)}{z} - \frac{1}{1+z}.$$

Proof. For $x = 0$, the result trivially holds. It then suffices to show that for all $x \in (0, \infty)$

$$h(x) := \frac{\ln(1+x)}{x} - b(z) - a(z)x \geq 0.$$

Note that $h(z) = 0$ and $h'(x) = a(x) - a(z)$. It follows from $\ln(1+x) > 2x/(x+2)$ for $x > 0$ (see (3) in [53]) that

$$\begin{aligned} a'(x) &= \frac{2\ln(1+x)}{x^3} - \frac{2}{x^2(1+x)} - \frac{1}{x(1+x)^2} \\ &\geq \frac{4}{x^2(2+x)} - \frac{2}{x^2(1+x)} - \frac{1}{x(1+x)^2} \\ &= \frac{1}{(2+x)(1+x)^2} > 0, \end{aligned}$$

thus $a(x)$ is an increasing function.

We separate two cases: for $x \in (0, z]$, we have $a(x) \leq a(z)$ and $h'(x) \leq 0$, thus $h(x)$ is decreasing on $(0, z]$ and $h(x) \geq h(z) = 0$; for $x \in [z, \infty)$, we have $h'(x) = a(x) - a(z) \geq 0$, $h(x)$ is increasing on $[z, \infty)$ and $h(x) \geq h(z) = 0$. We have shown that $h(x) \geq 0$ for $x > 0$. □

Note that the bound is tight at both $x = 0$ and $x = z$. The lemma does not cover the case $z = 0$, however, at the limit, we get $\lim_{z\to 0^+} b(z) = 1$ and $\lim_{z\to 0^+} a(z) = -1/2$, which corresponds to the bound given by Lemma 27.

Corollary 22. *Let $X \sim \mathrm{Bin}(n, p)$ and $\mu := E[X]$, then it holds for all $c > 0$ that*

$$E[\ln(1+cX)] \geq \ln(1+c\mu) - \frac{c}{2} \cdot \frac{c\mu}{1+c\mu}.$$

Proof. For $p = 0$ (or $\mu = 0$), the bound is trivial. Otherwise, for $p > 0$, applying Lemma 21 with $z = c\mu$ gives $\ln(1+cX) \geq b(c\mu)cX + a(c\mu)(cX)^2$, hence

$$\begin{aligned} &E[\ln(1+cX)] \\ &\geq b(c\mu)c\mu + a(c\mu)c^2\mu(1-p+\mu) \end{aligned}$$

$$
\begin{aligned}
&= \ln(1+c\mu) - c(1-p)\left(\frac{\ln(1+c\mu)}{c\mu} - \frac{1}{1+c\mu}\right) \\
&> \ln(1+c\mu) - c\left(\frac{1}{2}\cdot\frac{2+c\mu}{1+c\mu} - \frac{1}{1+c\mu}\right) \\
&= \ln(1+c\mu) - \frac{c}{2}\cdot\frac{c\mu}{1+c\mu}.
\end{aligned}
$$

The last inequality is due to $1-p<1$ and $\ln(1+x)/x < (1/2)(x+2)/(x+1)$ for $x>0$ (see (3) in [53]). □

The following lemma shows that all mask-based crossover operators satisfy (C3) with $\varepsilon = 1/2$ for Om and Lo functions.

Lemma 23. *If $x \sim p_{\mathrm{xor}}(u,v)$, where p_{xor} is a mask-based crossover, then:*

1. *If $\textsc{Lo}(u) = \textsc{Lo}(v) = j$, then $\Pr(\textsc{Lo}(x) \geq j) = 1$, otherwise $\Pr(\textsc{Lo}(x) > \min\{\textsc{Lo}(u), \textsc{Lo}(v)\}) \geq 1/2$.*
2. *$\Pr(\textsc{Om}(x) \geq \lceil(\textsc{Om}(u) + \textsc{Om}(v))/2\rceil) \geq 1/2$.*

Proof. 1) When $\textsc{Lo}(u) = \textsc{Lo}(v) = j$, in mask-based crossover operators, the two bitstrings x', x'' have j leading ones. So does the returned bitstring, i. e. with probability 1.

If $\textsc{Lo}(u) \neq \textsc{Lo}(v)$, we can assume w. l. o. g. that $\textsc{Lo}(v) = j$ and $\textsc{Lo}(u) > \textsc{Lo}(v)$. Then v has a 0 while u has a 1 at position $j+1$. So, one of the bitstrings x', x'' in the mask-based crossover will inherit the 1 at that position and the other will inherit the 0. This implies that one of them has fitness at least $j+1$ and with probability $1/2$ it is returned as output.

2) Each bit of u and v is copied either to x' or to x'', therefore $|x'|_1 + |x''|_1 = |u|_1 + |v|_1$, which means that $\max\{|x'|_1, |x''|_1\} \geq \lceil(|u|_1 + |v|_1)/2\rceil$. The output is chosen with probabilities $1/2$ to be copied either from x' or x'', and the result follows. □

The following two lemmas are used in the proof of Theorem 17.

Lemma 24. *If $X \sim \mathrm{Bin}(n,p)$, where $p>0$, then for all $y \in \mathbb{R}$ it holds that*

$$E[\max(X,y)] < \max(np, y) + (1/2)\sqrt{np}.$$

Proof. By Jensen's inequality w. r. t. the square root, we have

$$
\begin{aligned}
E[|X-y|] = E\left[\sqrt{(X-y)^2}\right] &\leq \sqrt{E[(X-y)^2]} \\
&= \sqrt{np(1-p) + (np-y)^2} \\
&\leq \sqrt{np(1-p)} + |np-y|,
\end{aligned}
$$

where the last inequality uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. Therefore, it holds that

$$
\begin{aligned}
E[\max(X,y)] &= E[(1/2)(X + y + |X-y|)] \\
&\leq (1/2)(np + y + |np-y| + \sqrt{np(1-p)}) \\
&< \max(np, y) + (1/2)\sqrt{np}.
\end{aligned}
$$

□

Lemma 25. *If $X \sim \mathrm{Bin}(n,p)$, where $p > 0$, then it holds that $E[X \mid X > 0] \leq np + 1$.*

Proof. By definition,

$$\begin{aligned} E[X \mid X > 0] &= \sum_{i=1}^{n} i \Pr(X = i \mid X > 0) \\ &= \frac{1}{\Pr(X > 0)} \sum_{i=1}^{n} i \Pr(X = i \cap X > 0) \\ &= \frac{1}{\Pr(X > 0)} \sum_{i=0}^{n} i \Pr(X = i) \\ &= \frac{E[X]}{\Pr(X > 0)} = \frac{np}{1-(1-p)^n} \leq np + 1, \end{aligned}$$

where the last inequality follows from Lemma 28. □

Appendix E Other results

The following well-known results are included for completeness.

Lemma 26 (Lemma 6 in [14])**.** *If $X \sim \mathrm{Bin}(\lambda, p)$ with $p \geq (i/\lambda)(1+\delta)$, then $E\left[e^{-\kappa X}\right] \leq e^{-\kappa i}$ for any $\kappa \in (0, \delta)$.*

Lemma 27 (Lemma 33 in [14])**.** *For all $x \geq 0$, $x \geq \ln(1+x) \geq x(1-x/2)$.*

Lemma 28 (Lemma 31 in [14])**.** *For $n \in \mathbb{N}$ and $x \geq 0$, we have $1-(1-x)^n \geq 1-e^{-xn} \geq \frac{xn}{1+xn}$.*

Lemma 29 (Lemma 3 in [15])**.** *For any $\varepsilon \in (0,1)$ and $\chi > 0$, if $n \geq (\chi+\varepsilon)\frac{\chi}{\varepsilon}$ then $(1-\varepsilon)e^{-\chi} \leq \left(1-\frac{\chi}{n}\right)^n \leq e^{-\chi}$.*

Lemma 30 (Corollary 3 in [13])**.** *Let $Y_1, \ldots, Y_n$ be n independent random variables with support in $[0,1]$ and finite expectations and $Y := \sum_{i=1}^{n} Y_i$. It holds for every $\delta > 0$ that*

$$\Pr(Y > E[Y] - \delta) \geq \min\left\{\frac{1}{13}, \frac{\delta}{1+\delta}\right\}.$$

References

[1] G. Ausiello and M. Protasi, "Local search, reducibility and approximability of NP-optimization problems," *Inform. Process. Lett.*, vol. 54, pp. 73–79, 1995.

[2] G. Badkobeh, P. K. Lehre, and D. Sudholt, "Unbiased black-box complexity of parallel search," in *Proc. of PPSN XIII*. Springer, 2014, pp. 892–901.

[3] H.-G. Beyer, H.-P. Schwefel, and I. Wegener, "How to analyse evolutionary algorithms," *Theor. Comput. Sci.*, vol. 287, pp. 101–130, 2002.

[4] T. Chen, J. He, G. Sun, G. Chen, and X. Yao, "A new approach for analyzing average time complexity of population-based evolutionary algorithms on unimodal problems," *IEEE Trans. Syst. Man. Cybern. B*, vol. 39, no. 5, pp. 1092–1106, 2009.

[5] T. Chen, P. K. Lehre, K. Tang, and X. Yao, "When is an estimation of distribution algorithm better than an evolutionary algorithm?" in *Proc. of CEC '09*. IEEE, 2009, pp. 1470–1477.

[6] T. Chen, K. Tang, G. Chen, and X. Yao, "On the analysis of average time complexity of estimation of distribution algorithms," in *Proc. of CEC '07*. IEEE, 2007, pp. 453–460.

[7] ——, "Rigorous time complexity analysis of univariate marginal distribution algorithm with margins," in *Proc. of CEC '09*. IEEE, 2009, pp. 2157–2164.

[8] ——, "Analysis of computational time of simple estimation of distribution algorithms," *IEEE Trans. Evol. Comput.*, vol. 14, no. 1, pp. 1–22, 2010.

[9] D. Corus, D.-C. Dang, A. V. Eremeev, and P. K. Lehre, "Level-based analysis of genetic algorithms and other search processes," in *Proc. of PPSN XIII*. Springer, 2014, pp. 912–921.

[10] D.-C. Dang, T. Friedrich, T. Koetzing, M. S. Krejca, P. K. Lehre, P. S. Oliveto, D. Sudholt, and A. M. Sutton, "Emergence of diversity and its benefits for crossover in genetic algorithms," in *Proc. of PPSN XIV*. Springer, 2016, pp. 890–900.

[11] D.-C. Dang, T. Jansen, and P. K. Lehre, "Populations can be essential in tracking dynamic optima," *Algorithmica*, vol. 78, no. 2, pp. 660–680, 2017.

[12] D.-C. Dang and P. K. Lehre, "Efficient optimisation of noisy fitness functions with population-based evolutionary algorithms," in *Proc. of FOGA XIII*. ACM, 2015, pp. 62–68.

[13] ——, "Simplified runtime analysis of estimation of distribution algorithms," in *Proc. of GECCO '15*. ACM, 2015, pp. 513–518.

[14] ——, "Runtime analysis of non-elitist populations: From classical optimisation to partial information," *Algorithmica*, vol. 75, no. 3, pp. 428–461, 2016.

[15] ——, "Self-adaptation of mutation rates in non-elitist populations," in *Proc. of PPSN XIV*. Springer, 2016, pp. 803–813, (arXiv:1606.05551).

[16] B. Doerr and C. Doerr, "Optimal parameter choices through self-adjustment: Applying the 1/5-th rule in discrete settings," in *Proc. of GECCO '15*. ACM, 2015, pp. 1335–1342.

[17] ——, "A tight runtime analysis of the (1+(λ, λ)) genetic algorithm on OneMax," in *Proc. of GECCO '15*. ACM, 2015, pp. 1423–1430.

[18] B. Doerr and M. Künnemann, "How the (1+λ) evolutionary algorithm optimizes linear functions," in *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '13. New York, NY, USA: ACM, 2013, pp. 1589–1596.

[19] S. Droste, "A rigorous analysis of the compact genetic algorithm for linear functions," *Nat. Comput.*, vol. 5, no. 3, pp. 257–283, 2006.

[20] S. Droste, T. Jansen, and I. Wegener, "On the analysis of the (1+1) Evolutionary Algorithm," *Theor. Comput. Sci.*, vol. 276, pp. 51–81, 2002.

[21] ——, "Upper and Lower Bounds for Randomized Search Heuristics in Black-Box Optimization," *Theor. Comput. Syst.*, vol. 39, no. 4, pp. 525–544, 2006.

[22] A. Eremeev, "Hitting times of local and global optima in genetic algorithms with very high selection pressure," *Yugosl. J. Oper. Res.*, vol. 27, no. 4, 2017, (To appear, arXiv:1606.05784).

[23] A. Eremeev and J. Kovalenko, "Optimal recombination in genetic algorithms for combinatorial optimization problems: Part I," *Yugosl. J. Oper. Res.*, vol. 24, no. 1, pp. 1–20, 2014.

[24] U. Feige, "On sums of independent random variables with unbounded variance, and estimating the average degree in a graph," in *Proc. of 36th STOC*, 2004, pp. 594–603.

[25] T. Friedrich, T. Kötzing, M. S. Krejca, and A. M. Sutton, "The benefit of recombination in noisy evolutionary search," in *Proc. of 26th ISAAC*. Springer, 2015, pp. 140–150.

[26] D. E. Goldberg, *Genetic Algorithms in search, optimization and machine learning*. Addison-Wesley, MA, USA, 1989.

[27] C. González, J. A. Lozano, and P. Larrañaga, "Analyzing the PBIL algorithm by means of discrete dynamical systems," *Complex Syst.*, vol. 12, pp. 465–479, 2000.

[28] B. Hajek, "Hitting-time and occupation-time bounds implied by drift analysis with applications," *Adv. Appl. Probab.*, vol. 14, no. 3, pp. 502–525, 1982.

[29] J. He and X. Yao, "Drift analysis and average time complexity of evolutionary algorithms," *Artif. Intell.*, vol. 127, no. 1, pp. 57–85, 2001.

[30] ——, "From an individual to a population: an analysis of the first hitting time of population-based evolutionary algorithms," *IEEE Trans. Evol. Comput.*, vol. 6, no. 5, pp. 495–511, 2002.

[31] T. Jansen and I. Wegener, "The analysis of evolutionary algorithms – a proof that crossover really can help," *Algorithmica*, vol. 34, no. 1, pp. 47–66, 2002.

[32] T. Kötzing, D. Sudholt, and M. Theile, "How crossover helps in pseudo-Boolean optimization," in *Proc. of GECCO '11*. ACM, 2011, pp. 989–996.

[33] M. S. Krejca and C. Witt, "Lower bounds on the run time of the univariate marginal distribution algorithm on onemax," in *Foundations of Genetic Algorithms (FOGA)*. ACM Press, 2017.

[34] P. Larrañaga and J. A. Lozano, Eds., *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*, ser. Genetic Algorithms and Evolutionary Computation. Springer, 2002, vol. 2.

[35] J. Lässig and D. Sudholt, "General upper bounds on the runtime of parallel evolutionary algorithms," *Evol. Comput.*, vol. 22, no. 3, pp. 405–437, 2014.

[36] P. K. Lehre, "Fitness-levels for non-elitist populations," in *Proc. of GECCO '11*. ACM, 2011, pp. 2075–2082.

[37] P. K. Lehre and P. T. H. Nguyen, "Improved Runtime Bounds for the Univariate Marginal Distribution Algorithm via Anti-Concentration," in *Proc. of GECCO '17*. ACM Press, 2017.

[38] P. K. Lehre and C. Witt, "Black-box search by unbiased variation," *Algorithmica*, pp. 1–20, 2012.

[39] P. K. Lehre and X. Yao, "Crossover can be constructive when computing unique input-output sequences," *Soft Comput.*, vol. 15, no. 9, pp. 1675–1687, 2011.

[40] ——, "On the impact of mutation-selection balance on the runtime of evolutionary algorithms," *IEEE Trans. Evol. Comput.*, vol. 16, no. 2, pp. 225–241, April 2012.

[41] A. Moraglio and D. Sudholt, "Principled design and runtime analysis of abstract convex evolutionary search," *Evol. Comput.*, 2015, posted Online.

[42] H. Mühlenbein and G. Paaß, "From recombination of genes to the estimation of distributions i. binary parameters," in *Proc. of PPSN IV*. Springer Berlin Heidelberg, 1996, pp. 178–187.

[43] H. Mühlenbein, "The equation for response to selection and its use for prediction," *Evolutionary Computation*, vol. 5, no. 3, pp. 303–346, 1997.

[44] F. Neumann, P. S. Oliveto, and C. Witt, "Theoretical analysis of fitness-proportional selection: landscapes and efficiency," in *Proc. of GECCO '09*. ACM, 2009, pp. 835–842.

[45] P. S. Oliveto and C. Witt, "Improved time complexity analysis of the simple genetic algorithm," *Theor. Comput. Sci.*, vol. 605, pp. 21–41, 2015.

[46] M. Pelikan, K. Sastry, and D. E. Goldberg, "Scalability of the bayesian optimization algorithm," *Intl Jour. on Approx. Reasoning*, vol. 31, no. 3, pp. 221–258, 2002.

[47] A. Prügel-Bennett, J. E. Rowe, and J. Shapiro, "Run-time analysis of population-based evolutionary algorithm in noisy environments," in *Proc. of FOGA XIII*. ACM, 2015, pp. 69–75.

[48] J. Rowe and D. Sudholt, "The choice of the offspring population size in the (1,λ) evolutionary algorithm," *Theor. Comput. Sci.*, vol. 545, pp. 20 – 38, 2014.

[49] J. Scharnow, K. Tinnefeld, and I. Wegener, "The analysis of evolutionary algorithms on sorting and shortest paths problems," *J. Math. Model. Algorithm.*, vol. 3, no. 4, pp. 349–366, 2004.

[50] J. L. Shapiro, "Drift and scaling in estimation of distribution algorithms," *Evol. Comput.*, vol. 13, no. 1, pp. 99–123, 2005.

[51] D. Sudholt, "How crossover speeds up building block assembly in genetic algorithms," *Evol. Comput.*, 2015, posted Online.

[52] D. Sudholt and C. Witt, "Update strength in EDAs and ACO: How to avoid genetic drift," in *Proc. of GECCO '16*, 2016, pp. 61–68.

[53] F. Topsøe, "Some bounds for the logarithmic function," in *Inequality Theory and Applications*, Y. J. Cho, J. K. Kim, and S. S. Dragomir, Eds. Nova Science Publishers, Incorporated, 2007, vol. 4, pp. 137–151.

[54] M. D. Vose, *The Simple Genetic Algorithm: Foundations and Theory.* MIT Press, Cambridge, MA, 1999.

[55] M. D. Vose and A. H. Wright, "Stability of vertex fixed points and applications," in *Proc. of FOGA III.* Morgan Kaufmann, 1995, pp. 103–113.

[56] I. Wegener, "Methods for the analysis of evolutionary algorithms on pseudo-Boolean functions," in *Evolutionary Optimization.* Springer US, 2002, vol. 48, pp. 349–369.

[57] D. Williams, *Probability with Martingales.* Cambridge University Press, 1991.

[58] C. Witt, "Runtime analysis of the (μ+1) EA on simple pseudo-Boolean functions," *Evol. Comput.*, vol. 14, no. 1, pp. 65–86, 2006.

[59] ——, "Upper Bounds on the Runtime of the Univariate Marginal Distribution Algorithm on OneMax," in *Proc. of GECCO '17.* ACM Press, 2017.

[60] Y. Yu, C. Qian, and Z.-H. Zhou, "Switch analysis for running time analysis of evolutionary algorithms," *IEEE Trans. Evol. Comput.*, vol. 19, no. 6, pp. 777–792, dec 2015.

[61] Q. Zhang and H. Mühlenbein, "On the convergence of a class of estimation of distribution algorithms," *IEEE Trans. Evol. Comput.*, vol. 8, no. 2, pp. 127–136, 2004.